Enabling New Speech Driven Services for Mobile Devices:

An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends

David Pearce

&
Chairman ETSI STQ-Aurora DSR Working Group

bdp003@email.mot.com

Motorola Limited
Jays Close
Viables Industrial Estate
Basingstoke
HANTS
RG22 4PD
UK

AVIOS 2000: The Speech Applications Conference, May 22-24, 2000 San Jose, CA, USA

Enabling New Speech Driven Services for Mobile Devices:

An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends

Abstract

In a distributed speech recognition (DSR) architecture the recogniser front-end is located in the terminal and is connected over a data network to a remote back-end recognition server. DSR provides particular benefits for applications for mobile devices such as improved recognition performance compared to using the voice channel and ubiquitous access from different networks with a guaranteed level of recognition performance. Because it uses the data channel, DSR facilitates the creation of an exciting new set of applications combining voice and data. To enable all these benefits in a wide market containing a variety of players including terminal manufactures, operators, server providers and recognition vendors, a standard for the front-end is needed to ensure compatibility between the terminal and the remote recogniser. The STQ-Aurora DSR Working Group within ETSI has been actively developing this standard and as a result of this work the first DSR standard was published by ETSI in February 2000. This paper presents an overview of the standard for the DSR Mel-Cepstrum front-end and compression algorithm together with its performance characteristics. The current activity in Aurora is to develop a future standard for an Advanced DSR front-end that will give half the error rate in noise compared to the Mel-Cepstrum.

1 Introduction to Distributed Speech Recognition

As has already happened in the wireline world, the trend to ever-increasing use of data communication is spreading to the mobile wireless world. As part of this, people want the ability to access information while on the move and the technologies to enable them to do this are now starting to be deployed. The small portable devices that will be used to access these data services cry out for improved user interfaces using speech input. At present, however, the complexity of medium and large vocabulary speech recognition systems are beyond the memory and computational resources of such devices.

Centralised servers can share the computational burden between users and enable the easy upgrade of technologies and services provided. Mobile voice networks, however, can degrade the performance obtained from centrally deployed recognisers. The degradations are a result of both the low bit rate speech coding and channel transmission errors. A Distributed Speech Recognition (DSR) system overcomes these problems by eliminating the speech channel and instead using an error protected data channel to send a parameterised representation of the speech, which is suitable for recognition. The processing is distributed between the terminal and the network. The terminal performs the feature parameter extraction, or the front-end of the speech recognition system. These features are transmitted over a data channel to a remote “back-end” recogniser. The end

result is that the transmission channel has minimal impact on the recognition system performance and channel invariability is achieved. This performance advantage is good for DSR services provided for a particular network e.g. by a network operator and additionally so for 3rd party DSR applications that may be accessed over a variety of different networks.

In summary the main benefits of DSR are as follows:

- Improved recognition performance over wireless channels.
  The use of DSR minimises impact of speech codec & channel errors that reduce the performance from recognisers accessed over digital mobile speech channels (see figure 4 for an example)
- Ease of integration of combined speech and data applications.
  Many new mobile multimodal applications are envisioned; such as the use of speech to access wireless internet content. The use of DSR enables these to operate over a single wireless data transport rather than having separate speech and data channels.
- Ubiquitous access with guaranteed recognition performance levels.
  There are currently 4 different major digital mobile systems each using several different speech codecs. These all produce different effects on recognition performance. Mismatches in the channels used for training and recognition can result in severe degradations in performance, while models that have been trained over many networks give a compromise performance. DSR on the other hand offers the promise of guaranteed level of recognition performance over every network. It uses the same front-end and there is no channel distortion coming from the speech codec and its behaviour in transmission errors.

2 DSR Standardisation in ETSI – (STQ-Aurora DSR working group)

To enable widespread applications using DSR in the market place, a standard for the front-end is needed to ensure compatibility between the terminal and the remote recogniser. The Aurora DSR Working Group within ETSI has been actively developing this standard over the last two years. To allow the optimisation of the details of the feature extraction algorithm a reference database and experimental framework has been established. The database is based on the original TIdigits database with controlled filtering and simulated noise addition over a range of signal to noise ratios from 20dB to –5dB. A reference recogniser configuration using Entropic's HTK HMM software was agreed to investigate changes solely in the front-end. This database has been made publicly available via the European Language Resources Association (ELRA). Extensive experimentation has been performed using this and other internal databases to agree on the feature extraction and to test the compression algorithms. The first DSR standard was published by ETSI in February 2000. This standard and some aspects of its performance are described in the sections that follow.

The Mel-Cepstrum was chosen for the first standard because of its widespread use throughout the speech recognition industry. It was acknowledged, however, that a front-

end that had improved performance in background noise would be desirable. The current activity of the Aurora working group is the development of a future standard for an Advanced DSR front-end that will give half the error rate in noise compared to the Mel-Cepstrum. A call for proposals has been publicised and in October 1999 eight organisations submitted candidates to the qualification phase. New evaluation databases with real-world noise are being developed to allow comparison between these algorithm candidates. A comprehensive set of evaluation criteria has been agreed and the proposers will make their final submissions to the selection phase in October 2000. The new standard is forecast to be published some time in 2002.

Even after the Advanced standard is produced it is expected that the Mel-Cepstrum standard will continue to be used, since noise robustness can also be achieved by algorithms used at the server back-end (eg adaptation and model compensation techniques).

It is anticipated that the DSR bitstream will be used as a payload in other higher level protocols when deployed in specific systems supporting DSR applications. Thus the standard does not cover the areas of data transmission or any higher level application protocols that may run over them. In this respect it is similar to speech codec standards where the codec is specified separately to the systems that use it.

It is expected that the DSR Front-end will be used in a variety of both current and future mobile networks and associated protocols. Even so it is beneficial to agree on the appropriate combinations of protocols in the chain from the client terminal device to the recognition server to support DSR applications. A new sub-group of Aurora called “DSR Applications and Protocols” is now being formed to address this. DSR has the power to improve the performance of applications like Voice Activated WAP pages, Voice Browsing (multi-modal I/O) or Large Directory Assistance. The sub-group will work in collaboration with other standards groups such as W3C (internet standards), WAP Applications and 3GPP to recommend how to integrate DSR into applications and extend existing protocols where necessary.

2 The DSR Mel-Cepstrum Front-end and Compression Standard

Figure 1 shows a block diagram of the processing stages for a DSR front-end. At the terminal the speech signal is sampled and parameterised using a Mel-Cepstrum algorithm to generate 12 cepstral coefficients together with C0 and a log energy parameter. These are then compressed to obtain a lower data rate for transmission. To be suitable for today’s wireless networks a data rate of 4800 b/s was chosen as the requirement. The compressed parameters are formatted into a defined bitstream for transmission.

The defined bitstream is sent over a wireless or wireline transmission link to the remote server where parameters received with transmission errors are detected and the front-end parameters are decompressed to reconstitute the DSR Mel-Cepstrum features. These are passed to the recognition decoder residing on the central server. The recogniser back-end is not part of the standard.

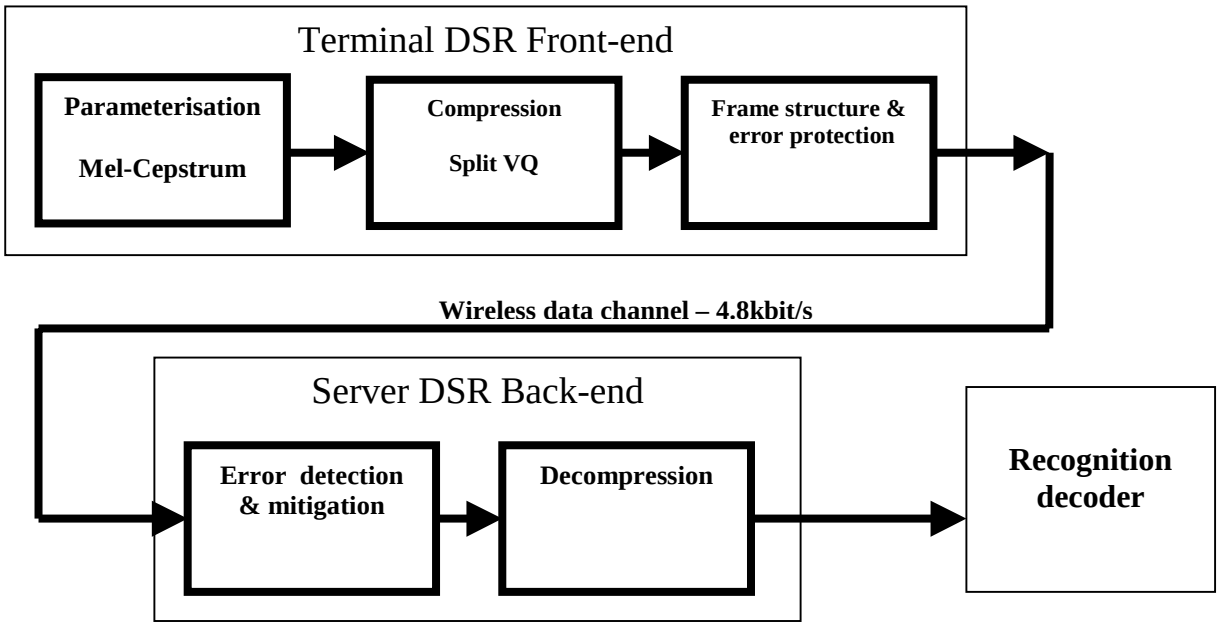
Since the data channels used for the transport of the DSR bitstream may be subject to errors (transparent data channels), special attention has been given to make the whole system robust to the types of burst errors that occur on wireless channels. To achieve this, error detection bits are added in the terminal DSR encoder as part of the bitstream and a special error mitigation algorithm is used at the decoder.

When developing the standard the following requirements were met:

- Mel-Cepstrum feature set consisting of 12 cepstral coefficients logE and C0
- Data transmission rate of 4800 b/s
- Low computational and memory requirements for implementation in the mobile terminals
- Low latency
- Robustness to transmission errors

Full details of the algorithms are given in the standards document [1] and can be accessed from the ETSI standards web site [see further information].

Figure 1: Block diagram of DSR system



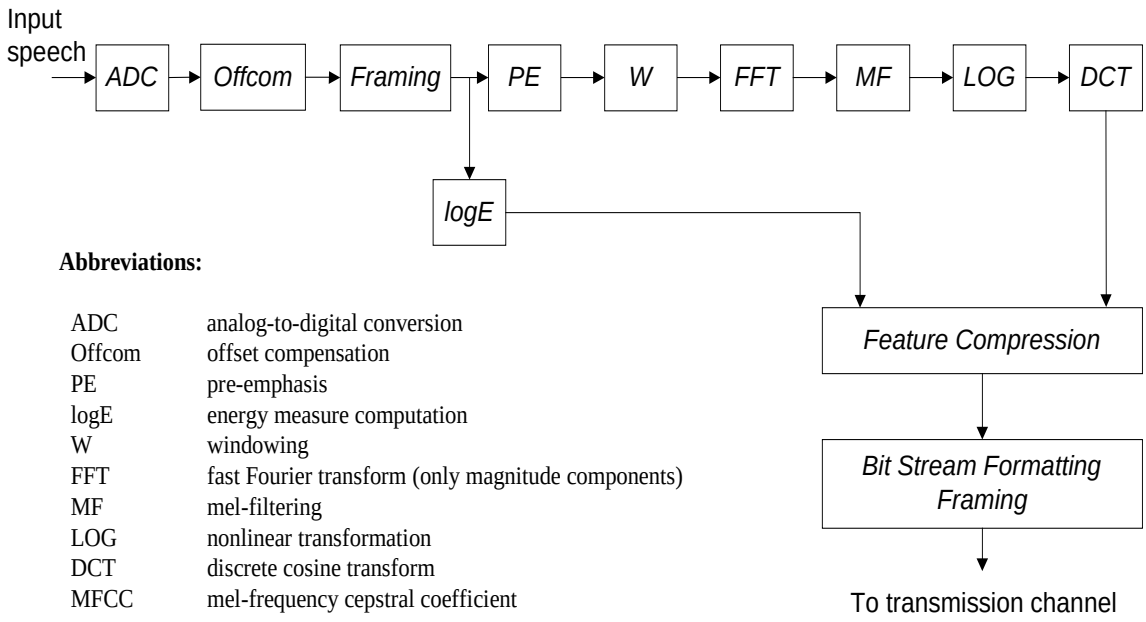
3 Optimisation of the Mel-Cepstrum Parameterisation

Nokia submitted the original proposal for the Mel-Cepstrum parameterisation to be used in the standard. The feature vector consists of 14 components composed of 12 cepstral coefficients (C1 to C12) together with C0 and a log energy parameter. C0 was included to

support algorithms that might need it in the back-end, such as noise adaptation (e.g. PMC). The basic building blocks of aspects of the Mel-Cepstrum will be very familiar to speech recognition experts using this front-end and the block diagram of the standard is shown in figure 2. There were, however, many detailed points concerning its implementation that needed to be agreed and there was extensive review and optimisation within Aurora. In particular the following aspects were tested and the conclusion reached for each is noted:

- Frame rate — 10ms
- Frame length — 25ms
- Filterbank definition — As proposed
- Energy calculation — changed to before preemphasis
- DC removal — GSM high pass filter
- Liftering — Removed

Figure 2: Block diagram of Mel-Cepstrum DSR Front-end standard



4 Compression Algorithm and its Performance Evaluation

Two proposals were made for the compression algorithm and the one eventually selected for the standard was from Motorola.

A split vector quantisation (VQ) algorithm is used to obtain a final total data rate of 4800 bits per second of speech. A codebook of size 64 is used for each pair of cepstral coefficients from C1 to C12 and 256 vectors are used for C0 and energy. This results in 44 bits per speech frame. Since the parameters may be transmitted over error prone

channels error detection bits (4 bits of CRC for each pair of speech frames) have been applied to the compressed data. For transmission and decoding the compressed speech frames are grouped into multiframes corresponding to 240ms of speech. The format is such that the bits corresponding to two frames may be transmitted as soon as they are ready. This results in only 10 ms additional latency at the terminal.

For full details see the DSR standard document [1]

The compression algorithms were tested on both small and large vocabulary tasks. The two areas that are important in evaluation of the performance of the compression algorithm are:

- That it does not produce a degradation in recognition performance compared with using the floating point mel-cepstrum parameters directly
- That the performance holds up in the presence of transmission errors typical of those that occur on wireless data channels (see section 5)

4.1 Small vocabulary database

The experimental framework for evaluation of DSR proposals is described in [2]. The performance of the quantizer has been evaluated by measuring the performance both with and without quantization applied to the test set. Models were trained using logE as the energy measure on unquantized data. The results are summarised in table 1 and show that the quantizer does not introduce significant performance degradation. The performance from a DSR recognition system can be considered transparent to the compression scheme proposed.

Table 1: Summary of recognition performance on Aurora noisy TIdigits evaluation databases

Training database	Unquantized test	Quantized test
8kHz Multicondition	**85.56**	**85.40**
8kHz Clean	**68.80**	**68.27**
16kHz Multicondition	**83.59**	**84.00**
16kHz Clean	**62.95**	**62.16**

4.2 Large Vocabulary Databases

The quantizer has also been tested over a wider range of tasks and recogniser configurations. The experiments presented here test the quantizers on large vocabulary tasks using sub-word models. The quantiser has been tested both on the Resource Management (RM) task and ATIS. Results presented here are for RM where two modelling configurations were used:

1) Monophone sub-word models

The first configuration uses simple monophone modelling of 42 phones, each having 3 states and 5 mixtures per state. Monophone models for the ATIS3 task were trained using the RM1, ATIS2, and ATIS3 speaker independent training sets. RM monophone models, however, were trained on only the RM1 speaker independent training set. In experiments using monophone models, a bigram language model is used.

2) Context dependent triphone sub-word models

This is more indicative of a state-of-the-art modelling approach. Context-dependent triphone modelling was performed, resulting in 1708 triphones. Each triphone is represented by 3 states, 5 mixtures per state. Triphone models for both ATIS3 and RM were trained using the RM1, ATIS2, and ATIS3 speaker independent training sets. A trigram language model was also utilised instead of the simpler bigram for these experiments.

The results are presented in the table 2:

Table 2: Quantizer evaluation using RM (Sept 92)

		8kHz		16kHz	
		No Quantizer	Motorola Quantizer	No Quantizer	Motorola Quantizer
Unquantized	Word %acc	89.88	89.96	91.60	91.99
Triphone	Word %corr	90.93	91.01	92.65	92.97
Trigram	Sent %corr	59.00	58.33	59.67	62.33
Unquantized	Word %acc	80.70	80.77	83.86	83.74
Monophone	Word %corr	83.31	83.00	85.62	85.46
Bigram	Sent %corr	40.00	42.00	40.67	40.67

5 Channel Error Robustness

The algorithm for error mitigation consists of two stages:

- Detection of speech frames received with errors
- Substitution of parameters when errors are detected

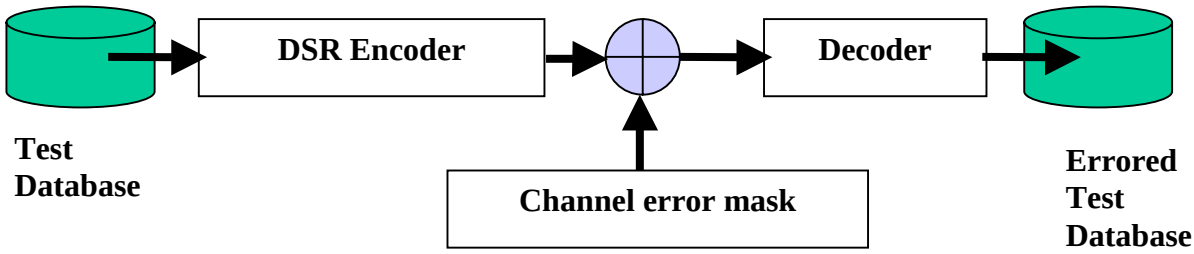
To detect the speech frames received with errors the 4 error detection bits on each pair of frames is used first. Since errors may be missed due to overloading of the CRC a heuristic algorithm that looks at the consistency of the parameters in the decoded frames is also used. It measures the difference between cepstral coefficients for adjacent frames and flags them as errored if it is greater than expected for speech. The thresholds used are

based on measurements of error free speech. If this algorithm was to run continuously then the number of misfirings would be too high, therefore it is only applied in presence of CRC errors.

When an frame is flagged as having errors then the whole frame is replaced with the a copy of the cepstral parameters for the nearest good frame received (occurring before or after the frame under consideration).

Robustness to channel errors has been measured according to procedures specified in Aurora on the noisy TIdigits task. Figure 3 shows the process for creating a test database that has been subject to channel error. The test set for these experiments is the digits at 20dB SNR with models trained on multicondition 8kHz data. In each of the channel experiments the encoded bitstream has had the bit error mask for the corresponding channel applied before decoding. The channels tested are for TETRA and GSM. The channel tested for TETRA was one at the edge of coverage while for GSM the 3 channels tested were those commonly used for GSM codec testing. EP1 represents a good quality channel, EP2 a medium one and EP3 a poor channel beyond the normal design target.

Figure 3: Evaluation tests for robustness to channel errors



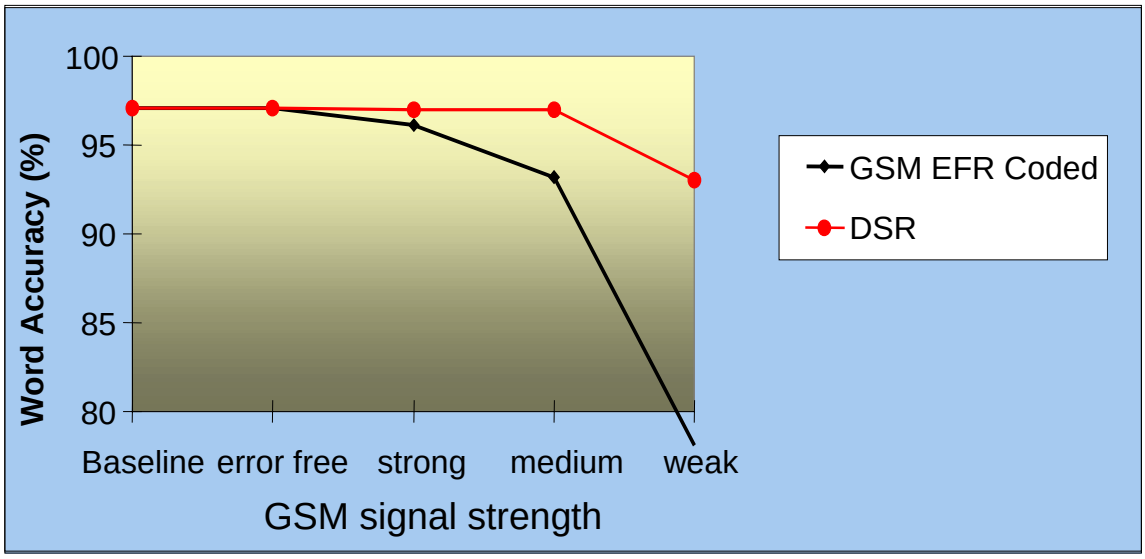
The results are presented in table 3 and show that for the TETRA channel at the edge of coverage and the EP1 and EP2 GSM data channels there is no significant degradation due to channel errors. For the GSM EP3 channel there is a 5% drop in performance. It should be noted that the EP3 channel is an extreme and represents an exceedingly poor channel. Speech audio quality for coded speech on such a channel is very poor. Speech recogniser performance based on speech coded with a GSM Enhanced Full Rate (EFR) codec and transmitted over an EP3 channel gives a performance of 78.1%. The significance of the performance benefits coming from DSR is illustrated in figure 4. For GSM channels the figure compares the error rates from DSR with those obtained using a GSM speech channel and the EFR codec.

The error robustness of the compression scheme is exceptionally good. Under most channel conditions within the designed coverage of a mobile system the degradation will be negligible.

Table 3: Error robustness performance over data channels with transmission errors

Channel	Noise 1	Noise 2	Noise 3	Noise 4	Average
Unquantized	97.26	96.19	97.61	98.33	**97.35**
error free quantize	97.42	96.28	97.76	98.33	**97.45**
TETRA TU50 20dB	97.17	96.01	97.26	98.18	**97.16**
GSM EP1	97.39	96.28	97.76	98.33	**97.44**
GSM EP2	97.39	96.28	97.64	98.24	**97.39**
GSM EP3	91.33	91.41	92.72	93.55	**92.25**

Figure 4: Performance with channel errors: DSR compared to a mobile speech channel



6 Latency

The total latency introduced over a wireless channel is made up of components from

- the encoder bitstream frame structure
- a minimum transmission time over the data channel (note other implementation specific delays are ignored here since they will vary from system to system but are independent of the bitstream frame structure)
- the decoder . This is the time from when a speech frame is received to when it can be made available to the back-end.

Quantized speech vectors are transmitted as frame pair packets corresponding to 2 frames. The additional latency introduced by the bitstream structure at the encoder is therefore 10ms (Note the Mel-Cepstrum FE parameterisation has a latency of 25ms while that for a frame pair is 25ms + 10ms = 35ms so the difference coming from the framing is 10ms).

There is a finite transmission time that depends on the data rate of the channel. In addition there is a delay as result of the frame structure of the data channel for the particular network (i.e. the interleaving in GSM) and implementation delays. For a particular channel though these will be constant and independent of the DSR bitstream. Once a frame packet pair of 92 bits has been processed and is ready for transmission the additional latency from the transmission of these parameters will be 92/data rate of channel (b/ms). For typical data channels such as GSM 9.6kbit/s this will be 9.6ms and for a TETRA 4.8kbit/s channel it will be 19.2ms.

The decoder can introduce an additional delay. This is the time taken from the receipt of the speech frame over the channel to when it is made available to the back-end recogniser. If an error free channel can be guaranteed then there would be no additional delay from the decoder (processing delay ignored). The error mitigation strategy used in the decoder makes use of two frame packet pairs. There is therefore an additional 20ms delay introduced by the decoder for error prone channels.

Encoder	**10ms**
Transmission [92/data rate of channel (b/ms)]	9.6ms GSM9.6 ~ 19.2ms TETRA4.8
Decoder - error free transmission	0ms
Decoder - with error mitigation	**20ms**
Total additional Latency	**30ms**

7 Terminal i/p characteristics

The goal of DSR is to have the best and consistent performance using server based recognisers. A further area where there is variability is in the terminal input characteristics including the frequency response of the microphone, analogue interface circuitry and the A/D converter. For GSM terminals the constraints on frequency response are specified in GSM 03.50 [3]. In Aurora an extensive set of experiments was conducted with data simulating the range different frequency responses conforming to this specification. A range of combinations of training and test conditions was tried. The conclusion was that if models were trained using data that had been filtered through G.712 and MIRS filtering (as specified by the ITU) and cepstral mean normalisation was used, then consistent performance was obtained for all filter responses conforming to 03.50. It was therefore concluded that the DSR standard was suitable for use in all terminals operating within the ranges of the characteristics as specified in GSM 03.50. DSR terminal developers should be aware that reduced recognition performance might be obtained if they operate outside the recommended tolerances.

8 Conclusions

The ETSI Aurora working group has completed the preparation of the standard for a DSR Mel-Cepstrum front-end and compression algorithm. The details of the Mel-Cepstrum feature extraction algorithm have been extensively tested and agreed. The compression algorithm produces no degradation in performance and the error robustness of the

compression scheme is exceptionally good. In addition it has low latency and the complexity is suitable for mobile handsets.

References

1. "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", April 2000.

2. "Experimental Framework for the Evaluation and Verification of Distributed Speech Recognition Front-ends": version 5, AU/134/98, 8th September 1998

3. GTS GSM 03.50:"Digital cellular telecommunications system (Phase 2+); Transmission planning aspects service in the GSM Public Land Mobile Network (PLMN) system (GSM 03.50)".

Further Information

The ETSI DSR standard document can be found at:

http://webapp.etsi.org/WorkProgram/Report_WorkItem.asp?WKI_ID=6400

For ETSI members more information about the Aurora working group is available from the ETSI FTP site where there are electronic copies of some of the documents from previous meetings is at

http://docbox.etsi.org/Tech-Org/STQ/Document/stq-aurora

To be able to access this area you need to be an ETSI member and obtain a password from them

http:/www.etsi.org

will provide you with information about how to do this.

Acknowledgement

My thanks to colleagues at Motorola who have contributed significantly to the compression part of this standard and its evaluation particularly Jon Gibbs, Jeff Meunier and Yan-Ming Cheng.

Thanks also to all the individuals who have contributed to this standard as participants in the Aurora Working Group. The main organisations represented there being Alcatel, British Telecommunications, Ericsson, France Telecom, Matra-Nortel, Motorola, Nokia, OGI, Qualcomm, Siemens, Sony, Texas Instruments.