Title: **Advanced DSR Front-end: Description of the Scope of the Standard**

Source: **David Pearce, Motorola**

Date: **15 March 2001**

Version: **1.0**

Summary

This document describes the component parts of a DSR system at the terminal and servers side and specifies which parts are within the scope of the front-end standard.

Description

Figures 1 & 2 show a block diagram showing the processing stages of a DSR system. These are split into the terminal (or client) side processing and the server side processing. Transmission between the client and server could be over either a wireless or wireline communication network or combination of the two.

The section numbers used below refer to the blocks in the diagrams.

Terminal/Client side DSR processing

1. Electro-acoustics

   This block refers to everything that occurs during the conversion of the sound pressure waveform to a digitised signal. These include the microphone transducer, analogue filtering, automatic gain control, analogue to digital conversion.

   The characteristics of the input audio parts of a DSR terminal will have an effect on the resulting recognition performance at the remote server. Developers of DSR speech recognition servers can assume that the DSR terminals will operate within the ranges of characteristics as specified in GSM 03.50 [1]. DSR terminal developers should be aware that reduced recognition performance may be obtained if they operate outside the recommended tolerances.

   GSM 03.50 will be referenced in the standard.

   Sampling frequencies of 8, 11 & 16 kHz are supported in the DSR standard.

   In the case of WI007 section 4.1 (“Introduction”) of the standard [2] comments on the input audio parts.

2. Speech detection or external control signal

In many applications a function performed at the terminal side will determine when the speech is to be processed and the DSR parameters transmitted over the network to the server.

This function may be performed for systems using circuit data transmission and is likely to be a commonly used component in services using transmission over packet data networks.

Three alternative ways in which this transmission control can be performed are:

- speech detection – the input speech signal is used to determine when there is speech activity
- push to talk – a user controlled button indicates when processing and transmission are to occur
- a signal coming from another software module

Speech detection is not part of the DSR front-end standard. The requirements for any speech detection algorithm to be used with DSR will be specified separately. (For example: The minimum number of frames that should to be sent ahead of the start of the speech utterance and after the end of the utterance is complete).

A recommendation can be made for the use of a particular speech activity detection algorithm that gives good results when used in conjunction with the advanced front-end standard, but it will not be normative.

3. Preprocessing

This block is optional and in most implementations it will be absent. It is not part of the DSR standard. Implementers may apply proprietary pre-processing stages ahead of the DSR standard. When doing so it is a manufacturer's responsibility to ensure that any pre-processing does not degrade performance of a DSR service. The result of any preprocessing should be to give a signal as if it had been recorded at a higher~~lower~~ signal to noise ratio. It should not result in spectral distortion or clipping of the speech signal. The output of this stage should remain within the constraints of GSM 03.50.

[Note that there will be issues to do with data collection from DSR terminals if different proprietary pre-processing stages are used. e.g. models trained with data collected from terminals using one proprietary preprocessing stage may not give good performance from terminals without any pre-processing or using an alternative. This point needs further discussion – options:

a) preprocessing is not allowed
b) the existence of a preprocessing stage could be flagged in the header.]

4. Parameterisation

   The frame based speech processing algorithm which generates the feature vector representation (B). This is specified in the front-end processing part of the DSR standard. In the case of WI007 it is the specification of the front-end feature vector extraction that produces the 14 element vector consisting of 13 cepstral coefficients and log Energy. See section 4.2 (“Front-end algorthm”) in the standard [2].

   After further processing stages the corresponding feature vector is recreated at the server (point C on figure 2).

5. Compression & Error protection

   The feature vector is compressed to reduce the data rate and error protection bits are added. This stage is specified as part of the DSR standard. In the example of WI007 the split vector quantisation algorithm is specified in section 5 of the standard (“feature compression algorithm”) and the error protection is defined in section 6 [2].

6. Formatting

   The compressed speech frames are formatted into a bitstream for transmission. Two types of data transmission will be supported:

   - Circuit data
   - Packet data

   For circuit data the format is defined as part of the DSR standard. In the case of WI007 this is described in section 6 of the standard (“Framing, bitstream formatting and error protection”). A multiframe format is defined with associated header and synchronization bits. For compatibility is expected that this same format will be used for WI008.

   For packet data transmission standardization will be via the IETF. DSR payloads for WI007 and WI008 will be defined for use in the Real Time Protocols (RTP).

Figure 1: Terminal/Client Side DSR Processing

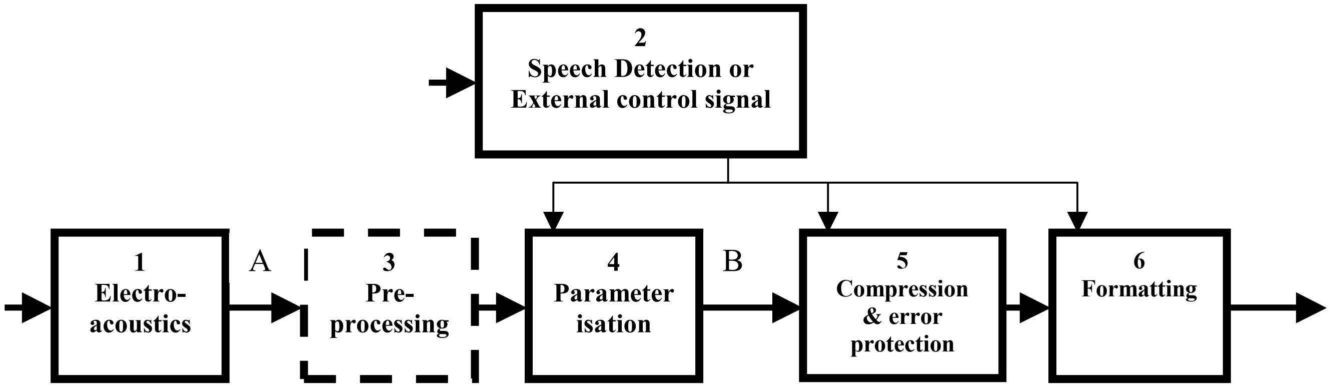
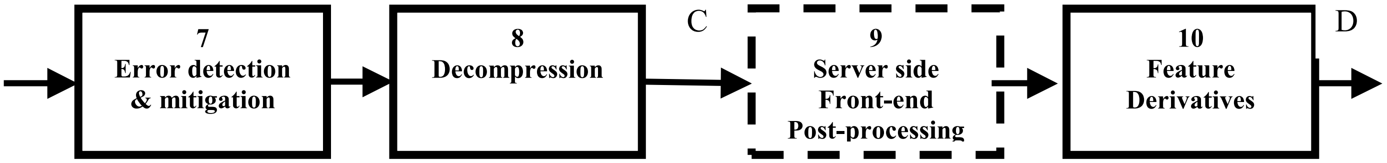


Figure 2: Server side DSR processing



Server Side DSR Processing

7. Error detection and mitigation

   This block specifies the method used to detect if there have been transmission errors and subsequent processing performed when these are detected. Its purpose is to minimize the effects of errors on the recognition performance. The algorithm used for error detection and mitigation is specified as part of the DSR standard.

Note that information derived from the feature vectors obtained after decompression (stage 8) may be used as part of the error mitigation.

In the case of the WI007 the error mitigation is specified in section 7.2.4 (“Error Mitigation”) of the standard [2].

8. Decompression

The compressed feature vectors are decompressed to give the feature vector corresponding that sent from the terminal at point B. (note that lossy compression is allowed and there will be an associated quantisation error)

Decompression forms part of the DSR standard.

9. Server side front-end post-processing

This block is optional. In some cases it may be appropriate to split the front-end processing that creates the front-end features between the terminal and the server. If further front-end processing is performed at the server side then it is specified as part of this block and will be presented as part of the DSR standard. In the case of WI007 there is no such processing performed.

10. Feature derivatives

It is common for speech recognition systems to make use of velocity and acceleration terms derived from the “static” feature vector.

Neither the number of parameters to be used by the recogniser or the method of calculation of the velocity or acceleration terms will not form part of the DSR standard. This processing is left open to be determined by the implementer of the recognition engine.

If a particular advantage can be demonstrated from a particular way of computing the derivative vector for the front-end selected for the standard, then this will be published as a recommendation in an annex to the standard but it will not be normative.

In the case of WI007 a method used by HTK to computing the derivatives has been used in experiments but no special method is recommended in the standard. For the experiments with HTK the first and second order derivatives of log energy and 12 cepstral coefficients were computed presenting a 39 element feature vector to the recognition engine at point D in figure 2.

References

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

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

Appendix 1

The following text is a copy of the statement describing the scope of standard as was agreed at the Aurora Amsterdam meeting 14th Feb 2001.

"What will be defined as part of the Advanced DSR front-end and compression standard consists of the terminal processing and server side processing that goes as far as the generation of the static features (in the example of the WI007 standard that means the 13 cepstral coeffs and logE).

The method and performance benefits of specific processing for dynamic features will be presented as an annex to the standard as a recommendation. (in the example of the WI007 standard that could mean velocity and acceleration terms)

Frame Deletion (FD) is a system level option for DSR that will be presented as an annex to the Front-end standard as a recommendation. A new activity in the working group will be started to address the requirements and recommendations for FD. The A&P subgroup will determine application requirements for FD than need to be supported."