Technical Specification Group Services and System Aspects **TSGS#21(03)0434**
Meeting #21, Frankfurt, Germany, 22-25 September 2003

3GPP TSG SA4 #27 S4-030543
Munich, Germany, 7-11 July 2003 Agenda Item: 9.1

Title: **Test and processing plan for default codec evaluation for speech enabled services (SES)**
Source: **TSG SA WG4 Codec**
Document for: **Information**
Agenda Item: **7.4.1**

Contact: **David Pearce, bdp003@motorola.com**
Version: 2.2

Summary

This document presents an update to version 2.0 of the test & processing plan for default codec evaluation for Speech Enabled Services previously approved at SA4 #26.

1. Introduction

Codec evaluation will be based on a framework which includes databases codecs and speech recognition engine. Evaluaters (as defined below) will be requested to use the same recognition engine for all codecs.

The following codecs have been submitted to the test:

1) AMR Codec and AMR WB Codec.
2) The ETSI DSR standard ES 202 050 for distributed speech recognition and its extension.

The evaluation framework for codec test is shown in Figure 1 and 2 below. Fig 1 applies for codecs with speech interface like a conventional speech codec and figure 2 applies for codecs with feature data interface like DSR optimised codecs.

The evaluation framework contains 2 processing stages:

- The candidate codec
- The speech recogniser from the evaluator

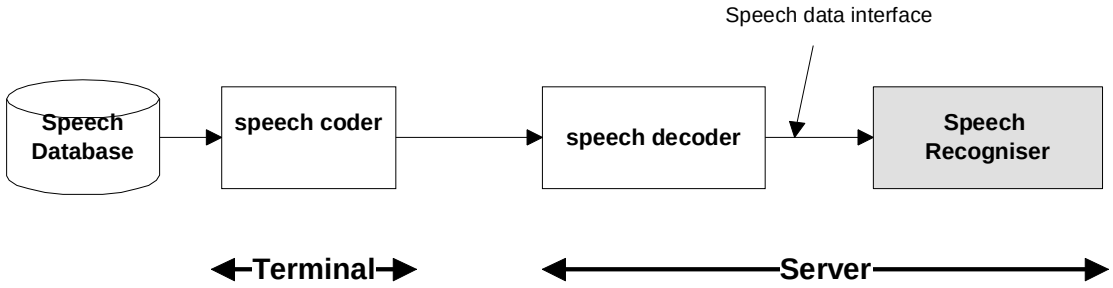

Figure 1: evaluation framework for speech codec (note that in this case the speech recognizer includes front-end and back-end decoder)

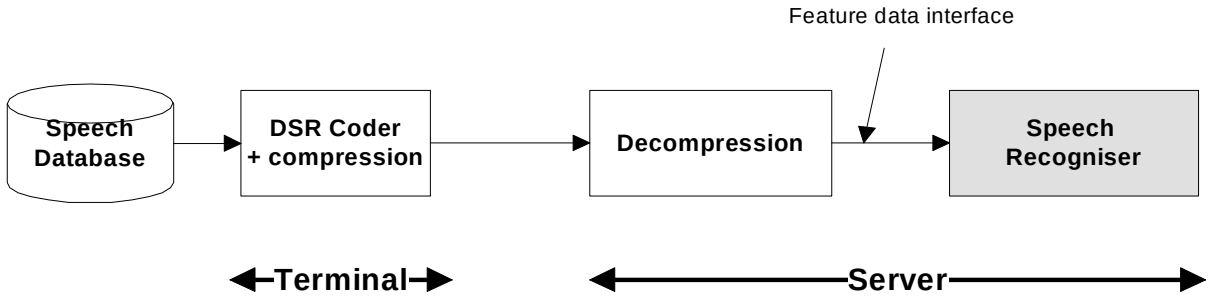

Figure 2: evaluation framework for DSR optimised codec (note that in this case the speech recogniser is back-end decoder only)

2. Recognition Engines

ASR vendors will perform the evaluations. Each ASR vendor will be provided with the database for the evaluation consisting of defined training and test sets (3GPP supplied databases). In addition ASR vendors proprietary databases will be used as well (ASR Vendor Supplied databases). Each ASR Vendor will run performance tests on these database considering both the AMR codec chain shown in figure 1 and the DSR optimised codec chain as shown in figure 2. ASR vendors have a free choice over the recogniser back-end configuration.

Evaluations will be performed using the bit exact implementations of the candidate codecs.

2.1 Recognizer for speech codecs based proposals

As AMR and AMR WB Codec can operate at several bitrates, a selection of bitrate has to be done for each test. Simulation of all AMR and AMR WB modes with all databases leads to practically unfeasible tests, therefore the number of Modes which are evaluated will be limited. For each selected bitrate the complete evaluation will be run on all databases. That means training and test is performed with that bitrate on the whole database. The following table shows the test conditions for AMR and AMR WB.

Bitrate	Codec	Sampling rate
4.75	AMR	8
12.2	AMR	8
12.65	AMR WB	16
23.85	AMR WB	16

Table 1: Test conditions for AMR and AMR WB Codec

2.1.1 Training & Testing

The training will be done using the coded & decoded speech data processed at the tested AMR bit rates as shown in the table above.

After speech decoder, any speech signal processing, e.g. compensating the coding artefacts or calculating the tonal language parameters, can be applied to the speech signal before calculating the actual recognition features.

2.2 Recognizer for DSR

Figure 2 shows the processing chain for a DSR front-end. The Advanced DSR Front-end (AFE) can operate with 8 or16kHz sampling rates. The feature extraction produces 12 mel-cepstral features (C1-C12), the zeroth order cepstral feature (C0) and log energy parameter (logE) at a 10ms frame rate. Recognisers may make use of either C0 or logE or both. The feature extraction is described in the ETSI standard document for ES 202 050 [1]. The static feature vector may be subject to further processing of the evaluators choice to produce dynamic features. The software for the DSR standard contains an example implementing the recommended way of derivative calculation although evaluators are free to use their own alternatives.

In addition to the cepstral features the DSR AFE extension provides a pitch feature that may optionally be used as a feature to assist recognition when processing tonal languages. The raw pitch feature may be subject to further processing of the evaluators choice to produce tonal features to supplement the cepstral feature vector (e.g. smoothing or derivative calculation).

2.2.1 Training &Testing

Training should be performed with the features after compression and decompression with an error free channel. The same feature post-processing should be used for training as for recognition.

2.3 Usage of VAD for frame dropping

For the purpose of these performance evaluations no voice activity detector will be used for frame dropping either for discontinuous transmission at the terminal or at the recognition engine at the server.

3 Codec Evaluations

3.1 Recognition experiments under error-free channel

Testing has been arranged to cover a range of tasks as shown in the list below:

1. Connected digit recognition task

   Aurora-2
   Aurora-3
   Vendor 2 In-car Japanese, German, US English
   Vendor 1 US English in-car
   Vendor 1 Mandarin Embedded corpus (digits)

2. Sub-word trained model recognition task

   Nokia Mandarin Chinese name dialling (tone recognition ignored in performance scoring)
   Vendor 2 In-car
   - Japanese,
   - German,
   - US English

   Vendor 1 Mandarin Embedded Corpus (names /street names /organization names/commands)
   Vendor 1 US English in car (commands, addresses, radio-controls, navigation, lifestyle information services and points-of-interest)

3. Tone confusability task

   Nokia Mandarin Chinese name dialling (tone recognition taken into account in performance scoring)

4. Channel error task.

   Aurora-3 Italian

Database Source	**Database**	**Evaluator**
3GPP supplied	Aurora-2	Vendor 2
	Aurora-3 German	Vendor 2
	Aurora-3 Spanish	Vendor 2
	Mandarin Name Dial	Vendor 1
	Aurora-2	Vendor 1
	Aurora-3 Spanish	Vendor 1
	Aurora-3 Italian	Vendor 1
ASR Vendor supplied	Mandarin Embedded PDA	Vendor 1
	US English In-Car	Vendor 1
	US English In-Car	Vendor 2
	German In-Car	Vendor 2
	Japanese In-Car	Vendor 2

Table 2: Table of databases for 8kHz Evaluations

Database Source	Database	Evaluator
3GPP Supplied		
	Aurora-3 Spanish	Vendor 2
	Aurora-3 Spanish	Vendor 1
	Aurora-3 Italian	Vendor 1
ASR Vendor Supplied	Mandarin Embedded PDA	Vendor 1
	US English In-Car	Vendor 1
	US English In-Car	Vendor 2
	German In-Car	Vendor 2
	Japanese In-Car	Vendor 2

Table 3: Table of databases for 16kHz Evaluations

3.2 Recognition experiments under channel errors

For the purposes of testing under channel errors the Aurora-3 Italian database with the well-matched training and testing condition will be used.

Each codec will be tested under error free channel and with average channel BLERs of 1%, 3%.

Recognition tests will be conducted by SpeechWorks and IBM using the supplied test sets. Models for these tests will be trained on the error free training data.

Codec for SES will be used with PSS over UTRAN, EGPRS and GPRS channels.

EGPRS (/GPRS) channel:

Simulations for GPSR and EGPRS will be combined as coding schemes for CS1 ..CS4 and MCS1 .. MCS4 are equivalent. Thereby consideration of EGPRS channel is sufficient.
The following parameters will be used:

- Typical Urban condition
- Scenarios: pedestrian with 3 km/h speed
- no FH
- unacknowledged mode
- One 20msec Frame per RTP/UDP Packet
- One RTP/UDP Packet per RLC/MAC Block

2 BLER patterns for EGPRS will be provided namely EG_EP1 and EG_EP2
EG_EP1 = error condition in very good channel (mean BLER ~ 1 %)
EG_EP2 = error condition in good channel.(mean BLER ~ 3 %)

UTRAN Channel:

Error situation for UTRAN channel will be better (fast power control) than in EGPRS channel. The UTRAN channel is here approximated using the EG_EP1 error mask of the EGPRS channel.
Format of Error Pattern

Error Pattern will be provided which contain one Flag per Block indicating the error status of the block. An error insertion device is used to skip the frame if the flag equals TRUE. The error mask is applied to the aligned coded speech data. That means with the first speech file the error mask is read from the beginning, At the end of the speech file a pointer showing to the position in the error mask file is stored. When the next speech file is processed the error mask is read from the position the pointer refers to. This continues till the end of the error mask file is reached. Then the error mask file is rewinded and whole process starts again.

Error patterns will be applied to the test database for each candidate where one 20ms frame (corresponding to one frame per block) is deleted as indicted by the binary file. It is the responsibility of each party submitting a codec candidate for speech enabled services to provide the error insertion device and create the test database set for each channel and apply error mitigation as appropriate.
Each party should be able to show how error masks were applied and allow verification of test database if required by others.

8kHz

	Error Free	EG_EP1	EG_EP 2

DSR	X	X	X
AMR 4.75	X	X	X
AMR 12.2	X	X	X

16kHz

	Error Free	EG_EP 1	EG_EP 2
DSR	X	X	X
AMR-WB 12.65	X	X	X
AMR-WB 23.85	X	X	X

4. List of evaluators:

Test will be made by two ASR Vendors namely IBM and Speech Works acting as testlabs.

5. Cost of databases

Aurora-2	250 Euro
Aurora-3	1000 Euro per language
Mandarin database from High-Tech 863 program	6000 US Dollars

Appendix 1: Description of Evaluation Databases

A1. Introduction

Several databases are used for the evaluation framework. The composition of the databases considers the real world situation and the requirements of the recommendation criteria. Databases contain several languages including tonal languages for tonal confusion tests. The environmental conditions are considered by including databases with real world noise. The application requirements are considered by including several tasks like digit task and a name dialling task. The Databases are selected from both the former ETSI STQ Aurora databases, from additional proposals of SA4 member companies and proprietary databases proposed by ASR vendors.

Training and testing will be performed using the 16bit linear data as supplied in the database distribution or obtained by downsampling from 16kHz to 8kHz if appropriate for the test.

In the following sections a short description is provided for all the databases used.

A2. Aurora 2: Noisy TI Digits database

The original high quality TIDigits database has been prepared by downsampling to 8kHz, filtering with G712 (which has frequency response representative of GSM terminal characteristics) and the controlled addition of noise to cover a range of signal to noise ratios (clean, 20,15,10,5,0,-5dB) and 8 different noise conditions. The database consists of connected digit sequences for American English talkers and clean and multi-condition training sets are defined. A full description of the database and the test framework is given in reference [2].

There are 3 test sets; set A contains noises seen in the multi-condition training data, set B contains noises that have not been seen in the training data and set C uses M-IRS filtering and noise addition to test the combination of convolutional distortion and noise.

A3. Aurora 3: Multilingual Speechdat-Car Digits database

Over a period of 4 years the ETSI STQ-Aurora working group has developed a set of evaluation databases and test criteria. Their purpose has been to support the characterisation and selection of Distributed Speech Recognition (DSR) front-ends. The databases cover a range of environments (typical for mobile device users) and languages. These have been made publicly available and are widely used. More details are given are given in reports sited in the references. The databases and procedures have been used for the competitive selection of the Advanced DSR front-end standard ES 202 050 and is

summarised in references [11, 13]. For ETSI members further information is available at the ETSI Aurora web site [12].

Tests with Aurora 3 database allow to evaluate the performance of the codec on data that has been collected from speakers in a noisy environment. It tests the performance of the front-end with well matched training and testing as well as its performance in mismatched conditions as are likely to be encountered in deployed DSR systems. It also serves to test the front-end on a variety of languages: Finnish [1)], Italian, Spanish, German, and Danish [3,4,5,6,7]. It is a small vocabulary task consisting of the digits selected from a larger database collection called SpeechDat-Car. See reference [3] as an example of for descriptions of these databases for Finnish with baseline performances for the mfccFE. The databases each have 3 experiments consisting of training and test sets to measure performance with:

A) Well matched training and testing - Train & test with the hands-free microphone over the range of vehicle speeds so that the training and test sets cover similar range of noise conditions.

B) Moderate mismatch training and testing - Train on only of a subset of the range of noises present in the test set. For example, hands-free microphone for lower speed driving conditions for training and hands free microphone at higher vehicle speeds for testing.

C) High mismatch training and testing - Model training with speech from close talking microphone. Hands-free microphone at range of vehicle speeds for testing.

[1)] An consistency check of all Aurora 3 databases showed that SDC Finnish seems to have some problems. Therefore this database will not be considered [15].

A3.1. Distribution and Availability of Aurora Databases

All of the Aurora databases have been made available publicly through the European Language Distribution Agency ELRA [8].

Note: These databases are now widely accepted and used by the international speech research communities. Two special sessions on Noise Robustness have been organised at international conferences where the Aurora-2 and Aurora-3 databases have been used for the purposes of comparing the performance of different research algorithms. At EuroSpeech 2001 held in Aalborg, Denmark in Sept 2001, 20 papers were presented at the session and at ICSLP held in Denver, USA in Sept 2002, 29 papers were presented with results on these databases.

A4. Mandarin Chinese Database (proposal from Nokia)

Training database: Mandarin Chinese database from Chinese 863 High-Tech Program

Training set: 100 female and 100 male speakers. The database consists of 4 groups of different sentences; each group has 500-600 sentences approximately. Each speaker pronounces one group. The whole data is about 115 hours of speech.

Test database: Nokia Tonal language database

Test set: 10 male + 10 female speakers, 512 full name utterances per speaker, 124 different names in the vocabulary (two names differing only in tone count as different ones)

Test conditions are clean speech and speech with background noise.

A4.1. Distribution and Availability of Chinese Database([14])

Mandarin Chinese database which is used for training and is a public database collected by Chinese High-Tech 863 Program. Contact person Miss Xie Ying (yxie@htrdc.com, +86 10 68339172).
The test database is available from Nokia under NDA agreement exclusively for this standardisation.

A5. Vendor 2 proprietary database

These databases are recorded in car simultaneously from a far-field and a near-field microphone. The corpora include digit strings, commands, and names (for voice dialing). Evaluations will be conducted for three languages: US English, German, and Japanese.

A6. Vendor 1 proprietary database

A6.1. US English In-Car Corpus

The database is used for Vendor 1's research experiments in embedded speech recognition. The recordings were made in stationary (with engine and a/c on) and moving (30mph and 60mph) cars with AKG-Q400 microphones placed on the mirror and visor. The corpus includes digit strings, commands, names and general English text. The training corpus is balanced for gender, accents, and other variations and is comprised of a very large number of speakers.

The test set also includes a large collection of speakers recorded in stationary and moving cars with a AKG-Q400 microphone placed on the mirror. The test corpus covers seven different tasks, digit strings, commands, addresses, radio-controls, navigation, vindigo (lifestyle information services) and points of interest.

A6.2. Mandarin Embedded Corpus

The database is designed for Mandarin speech recognition on handheld devices. This corpus is balanced for gender and other variations and is comprised of a very large set of speakers. The tasks covered in the corpus include digit strings/names /street names /organization names/commands etc. The test corpus is very similar to the training corpus. All recordings are made with a Lucent SD1100 microphone embedded into a PDA made in a university dormitory under usual background noise conditions.

References

[1] ETSI standard ES 202 050 "Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm", Oct 2002
[2] H G Hirsch & D Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"; Paris, France, September 18-20, 2000
[3] AU/225/00 "Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation", Nokia, Jan 2000
[4] AU/237/00 "Description and baseline results for the SpeechDat-Car Italian Database", Alcatel, April 2000
[5] AU/271/00 "Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation: Description and Baseline Results", UPC, Nov 2000
[6] AU/273/00 "Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation", Texas Instruments, Dec 2001
[7] AU/378/01 "Danish SpeechDat-Car Digits Database for ETSI STQ-Aurora Advanced DSR", Aalborg University, Jan 2001.
[8] http://www.icp.inpg.fr/ELRA/home.html the ELRA home page
[9] deleted reference
[10] Recommendation Criteria for default codec for speech enabled services (SES) S4-030075, 3GPP TSG SA4 meeting #25
[11] AU/372/01 "Overview of Evaluation Criteria for Advanced DSR front-ends, Version 8", Motorola, Dec 2001
[12] ETSI STQ-Aurora document archive: http://docbox.etsi.org/STQ/stq-aurora
[13] David Pearce, "Developing the ETSI Aurora Advanced Distributed Speech Recognition Front-end & What Next?", IEEE Automatic Speech Recognition and Understanding Workshop; ASRU 2001, Madonna di Campiglio, Italy, Dec 2001
[14] S4-020755 3GPP TSG SA4 meeting #25
[15] S4-030110 "Evaluation of usability of Aurora 3 databases"3GPP TSG SA4 meeting #25bis
[16] S4-030114 Reply LS on Transmission Aspects for Speech Enabled Services (SES)