STQ Aurora DSR Working Group

Title: **Advanced DSR Front-end: Definition of required performance characteristics**

Source: **Motorola**

Date: **21st December 2001**

Version: **4.1**

Version 1: AU/308/01 original
Version 2: AU/325/01 not issued
Version 3: Updated to reflect set of changes agreed in series of teleconference calls
Version 4: New large vocab performance improvement metric. Update to channel error mitigation table. Addition of summary table.

1 Introduction

ETSI STQ Work Item 007 produced the published DSR standard front-end algorithm based on Mel-Cepstrum technology [1]. ETSI STQ WI008 seeks to standardise a more advanced algorithm capable of at least matching Mel-Cepstrum's performance with low levels of background noise and significantly improving performance in more demanding environments.

This document specifies the performance characteristics required to select an algorithm for the Advanced DSR Front-end and compression. It updates and supersedes the qualification and selection criteria presented in AU/191/99 [2] taking account of new evaluation databases and further refinement of the requirements. It also defines the criteria to be used for the selection of the proposal for the Advanced DSR Front-end standard.

2 General requirements

2.1 Range of languages

The advanced front-end (AFE) shall be suitable for use with all the major languages of the world. For any language tested, the AFE should give improved recognition performance compared to the Mel-Cepstrum DSR standard. For practical reasons of resources and database availability it is not possible to test this requirement for all languages, but the AFE will be tested on a range of European languages. The AFE should not contain algorithm components that would be expected to give poor performance in other languages.

2.2 Range of noise environments

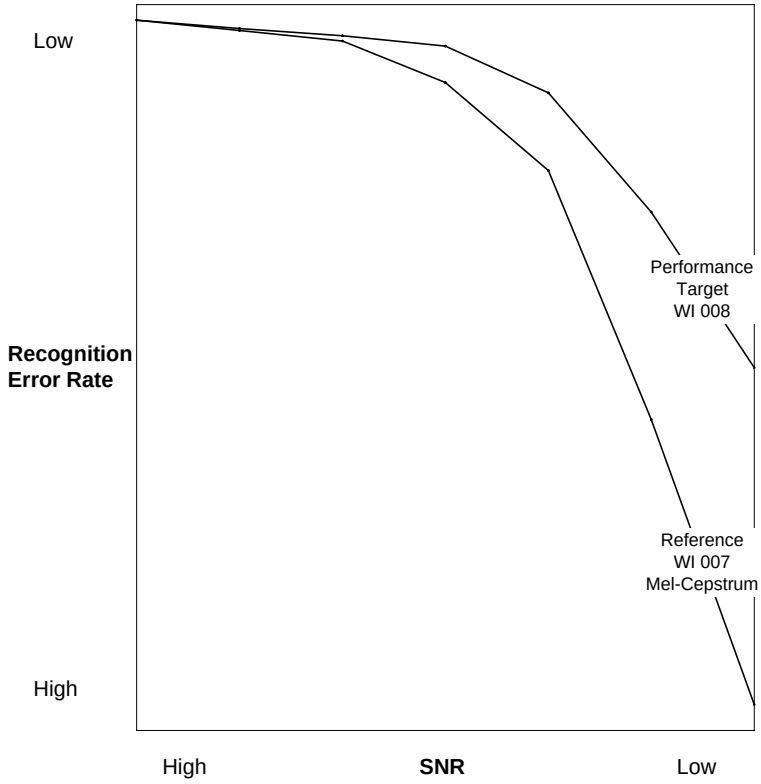
The AFE will be suitable for use in a range of background noises that are typical of the environments where mobile phones are used. For any noise environment tested the performance of the AFE will not be worse than that obtained from the Mel-Cepstrum standard.

2.3 Compatibility with back-end recognisers

The AFE will be suitable for use with recognisers based on Hidden Markov Model (HMM) technologies. It will be suitable for use with both whole-word and sub-word based HMM systems.

2.4 Improvement over Mel-Cepstrum DSR standard and graceful degradation in noise

The AFE will at least match the Mel-Cepstrum's performance with low levels of background noise and significantly improve performance in more demanding environments.



The figure below presents the recognition performance target in graphical format. It is expected that the advanced front-end algorithm will show graceful degradation in speech recognition performance as a function of degrading background noise conditions, similarly as shown by the reference WI007 algorithm [1].

Figure 1: The performance target for STQ WI008 Advanced Front-End standardisation.

3 Specific Requirements

3.1 Sampling Rates

Sampling rates of 8, 11 & 16kHz will be supported.

3.2 Speech Recognition Performance

1. The AFE must statistically match (or exceed) the performance of the reference WI007 Mel-Cepstrum algorithm with low levels of background noise. For the Aurora 2 database [5] the relevant test conditions are 'Clean' and '20dB SNR' for both clean and multicondition training. For the large vocabulary recognition task the relevant test is for clean training and testing at 8 and 16kHz.

Database	**Training**	**Testing**
Aurora 2	Clean	Clean, 20dB SNR
Aurora 2	Multicondition	Clean, 20dB SNR
Large Vocabulary 8kHz	Clean	Set 1
Large Vocabulary 16kHz	Clean	Set 1

1. The AFE must provide at least 25% improvement over the WI007 Mel-Cepstrum standard on small vocabulary recognition tasks under well-matched conditions at 8kHz sampling. For the Aurora 2 database this corresponds to the multi-condition training condition. For SpeechDat-Car this corresponds to the well-matched training and test set, averaged over the 5 languages.

1. The AFE must provide at least 50% improvement over the Mel-Cepstrum standard on small vocabulary recognition tasks under high mismatch conditions at 8kHz sampling rate. For the Aurora 2 database this corresponds to the average improvement with the clean training condition. For SpeechDat-Car this corresponds to the high-mismatch training and test set with the performance improvement averaged over the 5 languages.

1. The AFE must not show performance degradation relative to the Mel-Cepstrum in any of the 8 different noise conditions used in the Aurora 2 database at 8kHz sampling rate. For these purposes the performance for a particular noise condition is defined as the average over the SNRs from 20dB to 0 dB.

1. The AFE must provide at least 25% average improvement over the Mel-Cepstrum standard on large vocabulary recognition tasks with added background noise at 8kHz and 16kHz sampling rates for clean and multicondition training. The measure of the average improvement is using the recognition performance metric for the large vocabulary task described in section 5.2.

3.3 Complexity

The terminal side processing of the DSR front-end has to be able to be implemented within the resources of a typical mobile phone terminal. Accordingly the maximum complexity requirements for terminal side DSR front-end and compression have been taken to be those for the GSM AMR speech coding [8] (rounded up to the nearest integer).

Measure	**Requirement**

WMOPS	Less than 17
ROM size	Less than 15 kwords
RAM size	Less than 6 kwords

The definition of the wMOPS measure and recommendations on how to estimate the computation and memory requirements can be found in ETSI Technical document [7]. A word is defined as 16bits. These complexity measures are for the front-end feature extraction and compression and exclude the VAD.

3.4 Latency

The total additional front-end latency is defined as the time delay from the sampled speech in a frame at the terminal to the delivery of the corresponding complete feature vector to the recogniser at the server (excluding the transmission time). It includes the following components:

framing into analysis windows introduces nw/2 latency (where nw is the length of the window in ms)
algorithmic delay for the front-end features
feature compression at the terminal
decompression & channel error mitigation (Note that what is included is any inherent latency resulting from error mitigation scheme: Some error mitigation schemes may introduce a latency that is dependent on the channel error conditions. In these cases, what is included is the latency under zero channel errors that is inherent in the algorithm.)
post processing (i.e. the server side processing used to generate the full feature vector presented to recogniser from the received static parameters e.g. for dynamic features or alternatives)

The maximum total additional front-end latency is 220ms.

3.5 Data rate

The maximum permissible bitrate is 4.8kbit/s.

3.6 Feature Vector size

The maximum feature vector size to be presented to the recogniser after post processing (e.g. computation of derivative terms) is 60.

3.7 Compression

The combined process of compression and decompression should not result in a significant degradation in recognition performance.

For operational deployment a DSR system will include feature extraction and compression in combination. Performance of the advanced front-end will therefore be measured in this way and there is no separate requirement placed on the performance of the compression block alone. During performance evaluations model training will also be performed with compressed features.

3.8 Channel error resilience

The channel error resilience shall be equal or better than the WI007 Mel-Cepstrum standard in terms of absolute degradation in performance for EP2 and EP3 channel error masks. Channel error resilience will be measured for the small vocabulary tasks and not for the large vocabulary task.The specific tests are for Aurora-2 multi-condition training (average performance over all test sets and average performance at 20dB SNR) SpeechDat-Car Italian well-matched training and testing. In each case the performance degradation is measured as the drop in performance relative to the baseline with no channel errors and the error mitigation off. The performance requirement in terms of performance degradation with EP2 and EP3 channels is summarised in the following table.

Test	EP2	EP3
Aurora 2 multi-condition training - full test set	0.73 %	8.77 %
Aurora 2 multi-condition training – 20dB SNR test	0.59 %	6.86 %
SDC Italian well-matched	0.86 %	9.44 %

Note: The baseline performances for WI007 were obtained by applying channel error masks to the bitstream corresponding to the whole of each speech file. WI007 decoding and error mitigation is performed for all these frames. Only those frames in each file that fall within the ideal endpoints are used for the recognition tests. Model training is performed on error free compressed features with the ideal endpoints. See reference AU/377/01 [12] for details.

Since results may vary depending on the precise alignment of the error masks with the DSR payload an allowance of 0.5% (absolute) has been added to each baseline result obtained in WI007 when setting the requirement for the AFE.

4 Criteria for selection of proposal for the standard

The selection of the proposal that will become the Advanced DSR standard will be made in a single stage.

The criteria to be applied at the selection phase are as follows:

1) Any proposal not providing the information required for the selection phase will be dropped (these are specified in AU/372/01 [4]).
1) Any proposal not meeting the selection requirements for performance (on the small and large vocabulary evaluations), complexity, latency, channel error resilience and data rate as defined in section 3 of this document will be dropped.
1) The decision to select between proposals meeting all the requirements will be based on recognition performance. A single overall performance metric (defined below) that combines the scores for performance improvement relative to the Mel-Cepstrum DSR standard from the small and large databases will be used.

5 Recognition Performance Metrics

5.1 Recognition performance metric for small vocabulary recognition tasks

The small vocabulary databases used for AFE evaluations consist of:

1) Aurora 2 (Noisy TIdigits) with multi-condition and clean training sets and 3 test sets A, B & C. and
1) Aurora SpeechDat-Car subsets for 5 languages (Finnish, Italian, Spanish, German & Danish). For each language there are 3 training/test conditions (well matched, medium mismatch and high mismatch)

The following weightings are used to obtain an overall metric for the recognition performance combining the results from the different databases and test conditions.

Recognition metric (weightings %):

TIdigits 40
- A 40 B 40 C 20
- multicondition 50 clean 50

SDC 60 (equal weight to each of the 5 languages)
- Well-matched 40 medium-mismatch 35 high-mismatch 25

These weightings are also used to give a single measure of the average performance improvement compared to the Mel-Cepstrum standard. To compute this measure, the weightings are applied to the performance improvement (reduction in error rate) compared to the Mel-Cepstrum on the results for the individual databases.
i.e.

% improvement for Aurora 2 = 0.5 x (% improvement for multicondition training)
+ 0.5 x (% improvement for clean training)

where

% improvement for multicondition/clean training = 0.4 x (% improvement for set A)
+ 0.4 x (% improvement for set B)
+ 0.2 x (% improvement for set C)

and % improvement for set A/B/C = average % improvement for 20, 15, 10, 5 & 0dB SNRs.

% improvement for SDC = average % improvement for each language

where

% improvement for each language = 0.40 x (% improvement for well matched)
+ 0.35 x (% improvement for medium mismatch)
+ 0.25 x (% improvement for high mismatch)

Overall improvement = 0.4 x (% improvement for Aurora 2) + 0.6 x (% improvement for SDC)

5.2 Recognition performance metric for large vocabulary recognition tasks

Au33701 [9] describes the large vocabulary database based on controlled filtering and noise addition to the Wall Street Journal database (WSJ0). The specific evaluations using this database are described in AU***/01 [13]. The tests will produce 4 performance measures for the large vocabulary task.

8kHz clean training
8kHz multicondition training
16kHz clean training
16kHz multicondition training

There are 14 test sets for each experiment.

The performance result for each experiment is the average performance over the 12 of these

test sets with added noise 2-7, 9-14.

The performance improvement measure is the average improvement relative to the Mel-cepstrum across the test sets with added noise 2-7, 9-14. The measure of improvement used is specified in the equation below, which calculates the effectiveness of the front-end in reducing the gap in performance between clean and noisy conditions relative to the WI007.

Improvement = (WI007_evalX - AFE_evalX) / (WI007_evalX - WI007_cleantrain_eval1)

The overall performance metric for the large vocabulary tests is the average from the 4 experiments.

5.3 Overall recognition performance metric for small and large vocabulary recognition tasks

Overall metric = 0.2 large vocabulary metric + 0.8 small vocabulary metric.

5.4 Use of performance metrics

For the purpose of meeting the requirements for relative performance improvement compared to the Mel-Cepstrum the average performance improvement will be used.

For the purpose of proposal selection comparison will be made using the overall recognition performance metric.

5.5 Speech detection in performance evaluations

The baseline performances for the Mel-Cepstrum front-end will be measured using “ideal” endpoints. These endpoints are determined by recognition force alignment using the clean data files and the addition of 200ms at the start and end of each utterance. These endpoints are copied across to the corresponding noisy files. Baseline performances for the small vocabulary evaluations with endpoints are presented in AU/377/01 [12]. Baseline performances for the large vocabulary evaluations with endpoints are presented in AU/***/01 [13] (*in progress from ISIP*)

The performances from proposal submissions will be determined with a voice activity detection algorithm (VAD) of choice. The duration of the silence to add at the beginning or end of speech detection is a design choice for each proposal (the addition of 200ms at the start and end is only for the purposes of the baseline performances with the Mel-Cepstrum). The VAD must be suitable for on-line operation.

6 Summary Table of Performance Requirements Relative to WI007 Baseline

Requirement	Database	Training	Testing	Threshold
clean conditions	Aurora 2	Clean	Clean, 20dB SNR	> –1%
	Aurora 2	Multicondition	Clean, 20dB SNR	> –1%
	Large Vocabulary 8kHz	Clean	Set 1	> -2%
	Large Vocabulary 16kHz	Clean	Set 1	> -2%
small vocabulary well matched conditions	Aurora 2	multicondition	20dB to 0dB	> 25% improvement
	SDC 5 languages	Well matched WM 5 languages		> 25% improvement
small vocabulary missmatched conditions	Aurora 2	Clean	20dB to 0dB	> 50% improvement
	SDC 5 languages	High mismatch (HM) 5 languages		> 50% improvement
large vocabulary	Noisy WSJ	8 kHz clean train 8 kHz multicondition train 16kHz clean train 16 kHz multicondition train Average of test results improvement cf WI007 in test sets 2-7 & 9-14		> 25% improvement
Error mitigation (see section 3.8)	Aurora 2	multicondition	20 to 0dB	> -0.5% absolute
	Aurora 2	multicondition	20dB	> -0.5% absolute
	SDC Italian	Well matched		> -0.5% absolute

References

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

[2] “STQ WI008 Qualification and Selection Criteria; Version 1.6”, AU/191/99, Nokia, Aug 1999.

[3] “Advanced DSR Front-end: Description of the Scope of the Standard”, AU/307/01, Motorola, Mar 2001

[1] “Overview of Evaluation Criteria for Advanced DSR front-ends, Version 8”, AU/372/01, Chairman, Dec 2001

[2] “Second Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends”, AU/231/00, Ericsson & Motorola, March 2000

[1] ETSI SMG11 Tdoc SMG11 72/98, AMR Speech coding Development - Design Constraints (AMR-5), 23rd March 1998

[2] ETSI SMG11 AMR-9 "AMR permanent document (AMR-9) Complexity and delay assessment v1.0", 23rd March 1998

[1] ETSI SMG11 Tdoc SMG11 117/99, “Complexity verification report of the AMR codec, v2.0”, Alcatel, Philips, ST Microelectronics, Texas Instruments".

[2] AU/337/01 “Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task: Version 1.0”, Ericsson, June 2001

[1] AU/345/01 “Large Vocabulary Evaluation of Front-ends - Baseline Recognition System Description”, Mississippi State University, Aug 2001

[2] AU/374/01 “Small vocabulary evaluations: Baseline Mel-Cepstrum Performances with Speech Endpoints: Version 2”, Motorola, October 2001

[1] AU/377/01 “Performance of WI007 Mel-Cepstrum and Transmission over GSM Channel with Speech Endpoints”, Motorola, Nov 01.

[2] AU/***/01 “Large vocabulary evaluations: Baseline Mel-Cepstrum Performances” ISIP, ?? 2001