STQ Aurora DSR Working Group

Title:	**Advanced DSR Front-end: Definition of required performance characteristics**
Source:	**Motorola**
Date:	**18~~16~~ May~~March~~ 2001**
Version:	**2~~1~~.0**

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 pre-selection of proposals to go forward to the large vocabulary evaluations.

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].
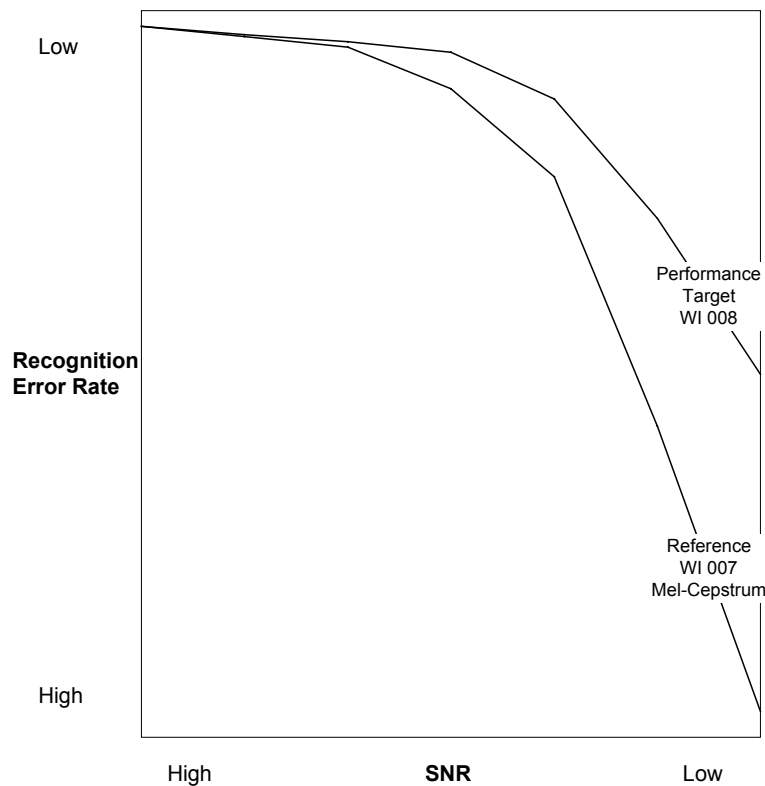
The performance target for STQ WI008 Advanced Front-End standardisation.

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 the large vocabulary recognition task the relevant test is for clean training and testing.

2. 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.

3. 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 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.

4. 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.

5. The AFE must provide at least 25% improvement over the Mel-Cepstrum standard on large vocabulary recognition tasks with added background noise at 8kHz and 16kHz sampling rates.

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.

3.4   Latency

The additional latency introduced by front-end and compression should not exceed 250 ms (Note that this figure excludes transmission time) but preference will be given to proposals achieving lower latencies. The additional latency is defined as the combination of front-end processing, compression and bitstream framing, occurring at the terminal equipment, together with the decoding at the DSR recognition server. It excludes the transmission time, which is dependent on the data channel.

The following latency figures should be determined.

> T-frameLength
> T-FrontEnd
> T-Compression+Framing
> T-Decoder
> T-Total

3.5   Data rate

(Previous statement) The maximum total data rate permissible from the bitstream is 9.6kbit/s but preference will be given to proposals achieving 4.8kbit/s and below.

(Suggested update) 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 computation of derivative terms is TBD ?60?. ~~(note that previously this was set to be about 200, but suggest we update it.)~~

3.7 Compression

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

For the Aurora 2 database, the absolute degradation in performance comparing testing with compressed or uncompressed data using models trained with multicondition uncompressed training shall be less than TBD ?1%?.

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 the small vocabulary testing this corresponds to the following measures:

Test	EP2	EP3
Aurora 2 multi-condition training - full test set	1%	8.4%
Aurora 2 multi-condition training – 20dB SNR test	1%	5.9%
SDC Italian well matched	1%	9.3%

4 Criteria for ~~pre-~~selection of proposals **for standard**~~to go forward to large vocabulary evaluations.~~

Based on the characteristics of each proposal submitted at the pre-selection round of evaluation a small number of proposals will be chosen to progress to the final round of selection. These proposals will undergo further testing on the large vocabulary evaluations and at 16kHz sampling.

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

1) Any proposal not providing the information required for the pre-selection phase will be dropped (these are specified in AU/275/00 [4]).
2) Any proposal not meeting the selection requirements for performance on the small vocabulary evaluations, complexity, latency, ~~compression,~~ channel error resilience and data rate defined in section 3 of this document will be dropped.
3) ~~The top proposal must achieve greater than TBD% ?40? reduction in error rate compared to the Mel-Cepstrum on the small vocabulary evaluations using the overall metric defined below. If this is not achieved then the AFE standardisation will be postponed.~~
4) ~~Any proposals passing criteria 1) and 2) above and obtaining more than (T - TBD ?5?)% will pass the pre-selection phase. Where T% is the overall performance metric of the top proposal.~~

Overall 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

2) 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 (both seen and unseen, equal weight to each language)
        well matched 40 medium 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 =   % 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)

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

[4] "Overview of Evaluation Criteria for Advanced DSR front-ends, Version 5", AU/275/00, Chairman, Dec 2000

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

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

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

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