DEVELOPING THE ETSI AURORA ADVANCED DISTRIBUTED SPEECH RECOGNITION FRONT-END & WHAT NEXT?

David Pearce
Motorola Labs & Chairman of Aurora
UKRL, Basingstoke, UK
bdp003@motorola.com

ABSTRACT

The ETSI STQ-Aurora DSR working group are developing the standard for the Advanced DSR front-end. One of the main goals of the advanced front-end is improved robustness to noise compared to the existing ETSI DSR standard for the Mel-Cepstrum front-end. The purpose of the paper is firstly to inform the wider speech research community about this activity and then to promote discussion on what further needs there are for DSR front-end standards. The scope of the DSR standard is described and the set of performance requirements that Aurora has specified for the Advanced Front-end. An important part of this the evaluation and characterisation of the performance of candidate front-ends on noisy databases and an overview of these is given. As the competition to select the best proposal draws to a close (submission deadline 28th Nov 2001) an interesting question is “what next?”.

1. SCOPE OF THE STANDARD

Figures 1 & 2 show a block diagram of 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 a wireline communication network or a combination.

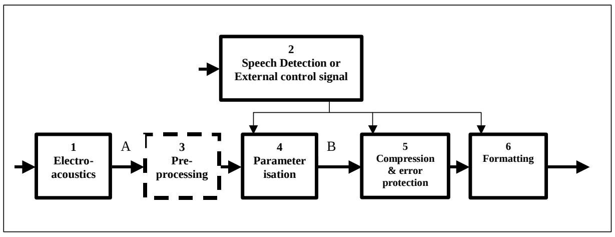

Figure 1: Terminal/Client Side DSR Processing

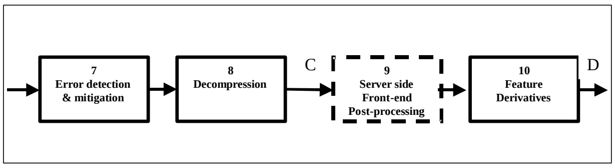

Figure 2: Server side DSR processing

Note that throughout the paper we will refer to the Mel-Cepstrum DSR standard as mfccFE and the Advanced DSR Front-end as AFE. The section numbers (after the “.”) used below correspond to the numbered blocks in the diagrams.

Terminal/Client side DSR processing

1.1 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 might be obtained if they operate outside the recommended tolerances.

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

1.2 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.

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 AFE standard, but it will not be mandated.

1.3 Pre-processing

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 pre-processing should be to give a signal as if it had been recorded at a higher 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.

1.4 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 mfccFE 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 algorithm") in the standard [1].

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

1.5 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 mfccFE 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 [1].

1.6 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 mfccFE 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 AFE.

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

2. EVALUATION DATABASES

2.1 Aurora 2: Noisy TI Digits – small vocabulary evaluation

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.

2.2 Aurora 3: Multilingual Speechdat-Car Digits – small vocabulary evaluation

The purpose of these tests is to evaluate the performance of the front-end on a database 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, Italian, Spanish, German, and Danish. It is a small vocabulary task consisting of the digits selected from a larger database collection called SpeechDat-Car. These experiments will be performed at 8kHz sampling rate. 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.

2.3 Aurora 4: Noisy WSJ – large vocabulary evaluation

AU/337/01 [4] describes the large vocabulary database based on controlled filtering and noise addition to the Wall Street Journal database (WSJ0). The tests will produce 4 performance measures for the large vocabulary task. The result in each case is the average performance improvement relative to the mfccFE baseline (at corresponding sampling rate) for the 14 test sets.

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

An HMM recogniser framework for this task has been prepared by the University of Mississippi for Aurora [5].

3. PERFORMANCE REQUIREMENTS

General requirements

3.1 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 mfccFE. 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.

3.2 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 mfccFE.

3.3 Compatibility with back-end recognisers

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

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

The AFE will at least match the mfccFE performance with low levels of background noise and significantly improve performance in more demanding environments.

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.

Specific Requirements

3.5 Sampling Rates

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

3.6 Speech Recognition Performance

The AFE must statistically match (or exceed) the performance of the reference mfccFE 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.
The AFE must provide at least 25% improvement over the mfccFE 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.

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.

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.

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.

3.7 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 (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 documents. A word is defined as 16bits.

3.8 Latency

(*Note that this requirement is still under discussion in Aurora*) The additional latency introduced by front-end and compression should not exceed 250 ms, 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 and post-processing at the DSR recognition server up to the point of presentation of the final feature vector to the recogniser (D in figure 2). It excludes the transmission time, which is dependent on the data channel.

3.9 Data rate

The maximum permissible bitrate is 4.8kbit/s (incl headers).

3.10 Feature Vector size

The maximum feature vector size to be presented to the recogniser after computation of derivative terms (or alternatives post-processing of static features) is 60.

3.11 Compression

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

In 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 HMM model training will also be performed with compressed features.

3.12 Channel error resilience

The channel error resilience shall be equal or better than the mfccFE 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. STANDARDISATION PROCESS

4.1 Qualification and selection phases

As is common practice for speech codec standardisation the process used to develop the AFE standard has been through an open competition to compare and select the best candidate proposal based on an agreed set of requirements. This has been done in two phases:

- **Qualification phase** - which was used to determine organisations wishing to submit candidates and whether any whether any of these were likely to meet the desired performance improvement in noise.
- **Selection phase** – submission of candidates with complete characterisation and documentation to determine whether the minimum performance requirements had been met for each candidate and to enable selection of the best proposal to be made.

Originally it was planned to have a pre-selection phase to narrow down the candidates to the top-performing cluster before a final selection phase involving more extensive evaluation. Due to some unexpected issues that arose during the at pre-selection phase (Jan 2001) concerning fair comparison of results it was decided to skip this phase and instead go directly to the final selection phase with all candidates completing the full set of evaluations.

4.2 Criteria for selection of proposals for the standard

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

- Any proposal not providing the information required for the selection phase will be dropped (these are specified in AU/275/00 [4]).
- 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.
- The decision to select between proposals meeting all the requirements will be based on recognition performance. A single overall performance metric that combines the scores for performance improvement relative to the mfccFE from the small and large databases is used.

5. WHAT NEXT?

By the time of the workshop final submissions for the advanced front-end will have been made (28th Nov 2001) and the winning proposal that will form the standard selected. So the interesting question is what next?

The need to extend the DSR front-end to allow for speech reconstruction and tonal language recognition has already been identified and a new ETSI work item has been created. What else will it be good to standardise? Here are some options to consider and we look forward to discussing it at the workshop:

- another front-end giving even better performance in noise? (the very advanced DSR front-end!)
- a fixed spectral representation as the common component of all front-ends while allowing flexible server side post processing variations?
- as above, but allowing a programmable spectral representation e.g. the number and spacing of filters?
- a speech codec that is good for recognition as well as speech transmission quality?

6. REFERENCES

[1] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.2 (2000-04), April 2000.

[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] Aurora document no. AU/225/00 “Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation”, Nokia, Jan 2000

[4] Aurora document no. 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

[5] Aurora document no. AU/345/01 “Large Vocabulary Evaluation of Front-ends - Baseline Recognition System Description”, Mississippi State University, Aug 2001

7. ACKNOLEDGEMENT

Many people participating in the Aurora working group have contributed to this work both directly and indirectly. There are too many people to list them all individually but honour them collectively for their creativity, time and energy given to progressing DSR standards.