STQ Aurora DSR Working Group:

Title : Overview of Evaluation Criteria for Advanced Distributed Speech Recognition Front-ends

Author : Motorola

Date : 16th October 2001

Version : 7

Revision history

Version 1: AU/187/00, Draft, 21 July 99
Version 2: Revised after Trento working group meeting
Version 3: Revised after Prague working group meeting
Version 4 (AU/265/00): Revised after Helsinki working group meeting
Version 5 (AU/275/00): Updated to include spreadsheet output and correct references.
Version 6 (AU/326/00): Not issued
Version 7 (AU/372/01): Updated in preparation for final submissions on 28th Nov 2001

1. Summary

This report provides an overview of the information to be provided for the evaluation of the performance of candidate Advanced DSR front-ends. It is intended to provide a guide and checklist for organisations conducting the work to characterize their proposed algorithm. Details of the individual tests are referenced elsewhere.

It is a requirement that proposals provide all the information specified below to be accepted. Incomplete proposals will be rejected.

2. Algorithm documentation

All proposals must provide documentation describing the details of all algorithms in the submission (front-end, compression and error mitigation).
The description should include a block diagram of the complete front-end algorithm and text and equations specifying the processing provided by each block.

Section 3.2 of the Mel-Cepstrum DSR standard provides an example of the level of detail required [1].

Candidates may optionally describe the particular advantages of the proposed algorithm.

The preferred format is Microsoft Word or pdf.

3. Software implementation

Candidates should supply source code for the implementation of their algorithm.

4 Recognition Performance

4.1 TIDigits with artificially added noise (Aurora 2 noisy TIDigits).

The original high quality TIDigits database has been prepared by downsampling and the controlled addition of noise to cover a range of signal to noise ratios (SNRs) and noise conditions. A full description of the database and the test framework is given in reference [2]. The database consists of connected digit sequences for American English talkers.

These experiments should be performed at the 8kHz sampling rate and with G712 filtering.

The multi-condition training set contains the speech data covering the range of signal to noise ratios (SNRs) of Clean, 20dB, 15dB, 10dB, and 5dB.

Testing is performed on a range of SNR conditions of clean, 20dB, 15dB, 10dB, 5dB, 0dB & -5dB An overall performance is obtained as the average for the 5 performances (expressed as word accuracy) between 0dB to 20 dB SNR.

4.2 Real-world noisy database (SpeechDat-Car subsets)

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 references [4,5,6,7] for descriptions of these databases for Italian, Finnish, Spanish & German. The databases each have 3 experiments consisting of training and test sets to measure performance with:

A) Well matched training and testing

Train on real-word data with matched microphone and coverage of a representative range of noise levels and types present in the test set.

B) Moderate mismatch training and testing

Train on range of SNRs consisting only of a subset of the range of noises (noise types and noise levels) present in the test set. 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.

4.3 Large Vocabulary Testing (noisy Wall Street Journal)

The purpose of the large vocabulary tests is to measure the performance of the front-end on a large vocabulary task with simulated noise addition. AU/337/01 describes the large vocabulary database based on controlled filtering and noise addition to the Wall Street Journal database (WSJ0). The recogniser system (developed and provided by the Institute for Signal and Information Processing at Mississippi State University) is typical of a state-of-the-art large vocabulary HMM sub-word system. Evaluation is performed at both 8kHz and 16kHz sampling rates. The large vocabulary task was chosen to evaluate the front-ends at 16kHz because it was expected that the greatest performance differences between 8 and 16kHz sampling would be observed. Clean and multicondition training sets are defined and the 14 test sets cover a range of noise types typical of mobile environments.

4.4 Recognition Performance Metrics

The recognition performance metrics are given in AU/371/01 [11]. Candidates should evaluate the performance of their proposal on the set of databases described above. All experiments are to be conducted with feature compression in both the training and testing and using a Voice Activity Detection (VAD) algorithm of choice. The spreadsheet AU/373/01 [12] should be used to submit detailed results and compute the overall performance metrics.

5 Compression

The degradation of feature compression is not measured separately. All results presented include compression in both the training and recognition. No additional information is needed.

6 Resilience to channel errors

The DSR front-end may be used on error prone channels. The purpose of this test is to measure the performance with channel transmission errors. Three simulated GSM channels are used EP1 (10dB C/I), EP2 (7dB C/I) and EP3 (4dB C/I). The channel error masks are those for GSM 9.6kbit transparent circuit mode data. The test framework used is that for the Aurora 2 noisy TIdigits and well-matched condition for the SpeechDat-Car databases for Italian as described in section 2. Models will be trained on the multi-condition, G712 filtered speech at 8kHz training portion of the database using acompressed parameterisation. Reference performance without error mitigation should be presented (ie error free conditions and error mitigation switched off). The performance of the error mitigation algorithm will be tested in the error free conditions (but with error mitigation on) and for the 3 error masks applied. Details of the method for alignment of the error masks with the speech data are described in [8]. Results will be presented as the word error rate for each channel condition and the absolute % fall in performance relative to testing with compression alone (i.e. error free channel and error mitigation off).

7 Data rate

Specify the amount of data needed to transmit the front-end parameters representing 1 second of speech in bits/s. The data rate should include headers.

8 Implementation complexity and delay

8.1 Computation

Evaluate the computation requirements in terms of wMOPS separately for front-end feature extraction and compression. The definition of the wMOPS measure and recommendations on how to estimate the computation and memory requirements can be found in ETSI Technical document [10]. Proponents are allowed to use a floating point ANSI C source code. It is the candidate's responsibility to correctly assess the complexity figures of an equivalent fixed-point implementation.

As well as presenting the figure for the total wMOPs, candidates should also show details of how this assessment was made. This is best be done by showing a breakdown of the software into its component module hierarchy and software loops. Note that in the situation where computation is signal dependent then the wMOPs figure presented should be for the theoretical worst-case situation.

Separate figures should be presented for the terminal and server components of the processing.

8.2 Memory

Evaluate the memory requirements separately for front-end feature extraction and compression. Memory should be expressed in words where a word is defined as 16bits. The maximum RAM and total ROM (excluding program ROM) requirements should be determined.

As well as presenting the figure for the total ROM & RAM, candidates should also show details of how this assessment was made.

Separate figures should be presented for the terminal and server components of the processing. At the terminal separate figures should be presented for both the front-end processing and the compression. At the server separate figures should be presented for the decompression and feature vector generation. Feature vector generation includes and processing subsequent to the decompression and prior to the presentation of the feature vector to the recogniser (it therefore includes computation of velocity or acceleration components).

8.3 Latency

Specify the total additional front-end latency as defined in section 3.4 of AU/371/01 [11].

The following figures should be presented in ms.

T-half framelength
T-FrontEnd algorithmic delay
T-Compression+Framing
T-Decoder & error mitigation
T-Post processing

T-Total

9 Feature vector size presented to the recognition server

Specify the feature vector size that the front-end will present to the server recogniser.

10 Format for presentation of results

Appendix 1 gives the format that should be used to present results. A spreadsheet in MS-Excel format is available [12] that will generate the summary sheet from the input data. The completed spreadsheet containing all the information above should be submitted for each proposal.

References

[1] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", April 2000
[2] AU/231/00 "Second Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends", Ericsson & Motorola, March 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 2000
[7] AU/233/00 "Requirements for Evaluation of Feature Compression presented as part of WI008 Front-end Proposals, Version 2", Nokia, March 2000
[8] AU/266/00 "Recognition with WI007 compression and transmission over GSM channel", Ericsson, Dec 00.
[9] ETSI SMG11 Tdoc SMG11 72/98, AMR Speech coding Development - Design Constraints (AMR-5), 23rd March 1998
[10] ETSI SMG11 AMR-9 "AMR permanent document (AMR-9) Complexity and delay assessment v1.4", Editor Frédéric Lejay (Alcatel), 23rd March 1998
[11] AU/371/01 "Advanced DSR Front-end: Definition of required performance characteristics, Version 3", Motorola, Oct 2001
[12] AU/373/01 "Front-end Evaluation Spreadsheet: Version 2" – zip file containing spreadsheets and documentation, Motorola, Nov 2001

Appendix 1: Spreadsheet Summary Sheet *(to be updated when new spreadsheet is complete)*

Advanced DSR Front-End Performance Characteristics Summary

Company / Submission Details

Recognition Performance

Noisy TI Digits (Aurora 2)

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition				
Clean Only				
Average				

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition				
Clean Only				
Average				

Overall recognition performance improvement:

SpeechDat-Car

Absolute performance						
Training Mode	Seen Databases			Unseen Databases		Average
	Italian	Finnish	Spanish	German	Danish	
Well Matched						
Medium Mismatch						
High Mismatch						
0.4W+0.35M+0.25H						

Performance relative to Mel-cepstrum						
Training Mode	Seen Databases			Unseen Databases		Average
	Italian	Finnish	Spanish	German	Danish	
Well Matched						
Medium Mismatch						
High Mismatch						
0.4W+0.35M+0.25H						

Compression Recognition Performance

Data Rate (Bit/s):

Absolute Performance

Uncompressed training, compressed testing				
Noisy TIDigits (Aurora 2) mulitcondition, uncompressed training				
	Set A	Set B	Set C	0.4A+0.4B+0.2C
Uncompressed				
Compressed				
Degradation				

SpeechDat-Car uncompressed training				
	Italian	Finnish	Spanish	Average
Uncompressed				
Compressed				
Degradation				

Models trained and tested with compressed features				
Noisy TIDigits (Aurora 2) mulitcondition, compressed training				
	Set A	Set B	Set C	0.4A+0.4B+0.2C
Uncompressed				
Compressed				
Degradation				

SpeechDat-Car Italian, compressed training				
	Well	Medium	High	0.4W+0.35M+0.25H
Uncompressed				
Compressed				
Degradation				

Performance Relative to Current Standard

Uncompressed training, compressed testing				
Noisy TIDigits (Aurora 2) mulitcondition, uncompressed training				
	Set A	Set B	Set C	0.4A+0.4B+0.2C
Uncompressed				
Compressed				

SpeechDat-Car uncompressed training				
	Italian	Finnish	Spanish	Average
Uncompressed				
Compressed			TBA	

Models trained and tested with compressed features				
Noisy TIDigits (Aurora 2) mulitcondition, compressed training				
	Set A	Set B	Set C	0.4A+0.4B+0.2C
Uncompressed				
Compressed				

SpeechDat-Car Italian, compressed training				
	Well	Medium	High	0.4W+0.35M+0.25H
Uncompressed				
Compressed				

Channel Error Resilience

Absolute Performance

	Noisy TI Digits (Aurora 2)				SpeechDat-Car Italian	
Condition	Results for 20 to 0 dB Multicondition		Results for 20dB SNR Test Only		SDC Italian Well Matched	
	Average %	Degradation	Average %	Degradation	Average %	Degradation
No Mitigation						
Error Free						
GSM EP1						
GSM EP2						
GSM EP3						

Performance Relative to Current Standard

	Noisy TI Digits (Aurora 2)		SpeechDat-Car Italian
Condition	Results for 20 to 0 dB Multicondition	Results for 20dB SNR Test Only	SDC Italian Well Matched
	Relative to current standard	Relative to current standard	Relative to current standard
No Mitigation			
Error Free			
GSM EP1			
GSM EP2			
GSM EP3			

Compexity and Latency

Complexity				
Measure	Terminal Feature Extraction	Terminal Compression	Server Decompression	Server Feature Vector Generation
CPU Load, wMOPs				
ROM, kwords				
RAM, kwords				

Total vector size presented to the recognition server:	

Latency	
Measure	Latency, ms
FrameLength	
FrontEnd	
Compression&Frame	
Decoder	
Total	