Technical Specification Group Services and System Aspects **TSGS#21(03)0440**
Meeting #21, Frankfurt, Germany, 22-25 September 2003

3GPP TSG SA4 #27 S4-030540
Munich, Germany, 7-11 July 2003 Agenda Item: SES

Title: **Recommendation Criteria for Default Codec for Speech Enabled Services (SES)**
Source: **TSG SA WG4 Codec**
Document for: **Approval**
Agenda Item: **7.4.3**

Contact: **David Pearce, bdp003@motorola.com**
Version: 2.0

Summary

This document provides the recommendation criteria for the default codec for speech enabled services (SES) as agreed at SQ SWG, SA4#27.

Updated to remove the 16kHz Mandarin Name dialling task and include agreed values for recommendations.

1. Introduction

This document defines recommendation criteria for the selection of the default codec for speech enabled services. These criteria are based on the design constrains [1] and performance evaluations described in the test and processing plan [2]. The recommendation is based on speech recognition performance and the details of the scoring system are described below.

2. Recognition performance

2.1 Overview

The set of databases used for the evaluations are defined in the Test and Processing Plan [2]. Each of these databases contains different types speech material covering a variety of tasks, environments and languages. Recommendation will be based on a score obtained from the recognition performance measured on each of these different databases. Section 2.3 describes how the scores from all the individual databases are combined using a weighting table (see also appendix 2).

2.2 Scoring on individual databases

For each database the reference performance is measured as the word error rate obtained from the ASR vendor’s system. This is the performance obtained from a state-of-the-art system from the ASR vendor assuming a transparent channel.

The performance (word error rate) on a given database is also measured with the ASR vendors system for a codec under test as described in the test and processing plan.

Scoring for tests performed with channel BLER described in section 3.1.2 of [2] will also be computed in a similar way. Note that only BLER of 1% and 3% are considered as part of the recommendation criteria.

2.3 Performance metric over all databases

The overall performance will be determined by averaging the absolute word error rate using the weightings presented in tables A2.1 for 8kHz sampling rate and A2.2 for 16kHz sampling rate of Appendix 2. The result of this weighted average is an overall measure of the average word error rate for each codec. This metric is called the “average word error rate”.

2.4 Comparisons between codecs

2.4.1 Low data-rate codec comparison

The two codecs under consideration at low data-rate are AMR 4.75 and DSR AFE with extension (5.6kbit/s). Only 8kHz sampling rate is considered since there is no AMR-WB codec at low data rate.

Table A2.1 in Appendix 2 shows the list of databases that will be tested and the weightings to be given to the scores obtained for each of these databases.

2.4.2 High data-rate codec comparison

At high data-rates the comparisons are made separately at 8kHz and 16kHz sampling rates.

2.4.2.1 8kHz sampling rate

The two codecs under consideration at high data-rate at 8kHz sampling are AMR 12.2 & DSR AFE and extension (5.6kbit/s).

Table A2.1 in Appendix 2 shows the list of databases that will be tested and the weightings to be given to the scores obtained for each of these databases.

2.4.2.2 16kHz sampling rate

The two codecs under consideration at high data-rate at 16kHz sampling are AMR-WB 12.65, & DSR AFE (5.6kbit/s).

Table A2.2 in Appendix 2 shows the list of databases that will be tested and the weightings to be given to the scores obtained for each of these databases.

3. Recommendation criteria

The recommendation procedure will consist of the following:

1. Candidates not compliant with all Design Constraints will be excluded from further consideration. (For the selection meeting, all candidates must provide justification document for meeting the Design Constraints.)

2. For the low data-rate comparison:
   - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 4.75kbps codec is more than 35% then the DSR codec and its extension will be recommended.
   - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 4.75kbps codec is less than 20% then the DSR codec will not be recommended.
   - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 4.75kbps codec is less than 20% then AMR will be recommended.
   - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 4.75kbps codec is between 20% and 35% then the performance results will be further considered by SA4 and if there is no consensus the results will be passed to SA for decision on what recommendation to make.

3. For the high data-rate comparison at 8kHz:
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 12.2kbps codec is more than 30% then the DSR codec and its extension will be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 12.2kbps codec is less than 20% then the DSR codec will not be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 12.2kbps codec is less than 20% then AMR will be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR 12.2kbps codec is between 20% and 30% then the performance results will be further considered by SA4 and if there is no consensus the results will be passed to SA for decision on what recommendation to make.

4. For the high data-rate comparison at 16kHz:
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR-WB codec is more than 25% then the DSR codec and its extension will be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR-WB codec is less than 15% then the DSR codec will not be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR-WB codec is less than 15% then AMR-WB will be recommended.
    - If the relative reduction in average word error rate for the DSR AFE codec and its extension compared to the AMR-WB codec is between 15% and 25% then the performance results will be further considered by SA4 and if there is no consensus the results will be passed to SA for decision on what recommendation to make.

References

[1] Design Constraints for default codec for speech enabled services (SES)
Tdoc S4-030248
3GPP TSG SA4 meeting #25bis, Berlin, Germany, 24-28 Feb 2003

[2] Test and Processing plan for default codec evaluation for speech enabled services (SES),
Tdoc S4-030395
3GPP TSG SA4 meeting #26, Paris, France, 5-9 May 2003

Appendix 1: Weighting scheme for results on each database

Each database in the test and processing plan [2] produces a set of results for different training conditions and test sets. The weighting scheme to be used to combine the different results to give a single average performance on each database is defined below

1. 3GPP supplied databases

1.1 Aurora 2

Database	Aurora 2		
Test Set	Set A	Set B	Set C
Weight of the test set	40 %	40 %	20 %

Table A1: Weighting scheme within the databases Aurora 2

Multicondition and clean trained results to be weighted equally.

2.2 Aurora 3

For the Aurora 3 databases there are three test sets, well matched, medium mismatch and high mismatch. These will be weighted equally.

2. ASR vendor supplied databases

Test sets within the ASR vendor supplied databases will be weighted equally.

Appendix 2: Weighting of evaluation databases

Task	Database	Evaluator	Task Weight	Database Weight
Digits	Aurora-3 German	Vendor 2	3/10	1/11
	Aurora-3 Spanish	Vendor 2		1/11
	Aurora-2	Vendor 2		1/11
	Aurora-3 Italian	Vendor 1		1/11
	Aurora-3 Spanish	Vendor 1		1/11
	Aurora-2	Vendor 1		1/11
	US English In-Car (digit test)	Vendor 2		1/11
	German In-Car (digit test)	Vendor 2		1/11
	Japanese In-Car (digit test)	Vendor 2		1/11
	US English In-Car (digit test)	Vendor 1		1/11
	Mandarin Embedded PDA (digit test set)	Vendor 1		1/11
subword	Mandarin Embedded PDA (names /street names /organization names/commands)	Vendor 1	4/10	1/6
	US English In-Car (commands, addresses, radio-controls, navigation, lifestyle information services and points-of-interest)	Vendor 1		1/6
	US English In-Car	Vendor 2		1/6
	German In-Car	Vendor 2		1/6
	Japanese In-Car	Vendor 2		1/6
	Mandarin Name dialling (baseform test)	Vendor 1		1/6
Tone confusability	Mandarin Name dialling (tone confusable test)	Vendor 1	1/10	1
Channel errors	1% BLER	Vendor 1	2/10	¼
	3% BLER	Vendor 1		¼
	1% BLER	Vendor 2		¼
	3% BLER	Vendor 2		¼

Table A2.1: Weighting of evaluation databases at 8kHz

Task	Database	Evaluator	Task Weight	Database Weight
Digits			3.5/10	
	Aurora-3 Spanish	Vendor 2		1/8
	Aurora-3 Italian	Vendor 1		1/8
	Aurora-3 Spanish	Vendor 1		1/8
	US English In-Car (digit test)	Vendor 2		1/8
	German In-Car (digit test)	Vendor 2		1/8
	Japanese In-Car (digit test)	Vendor 2		1/8
	US English In-Car (digit test)	Vendor 1		1/8
	Mandarin Embedded PDA (digit test set)	Vendor 1		1/8
subword	Mandarin Embedded PDA (names /street names /organization names/commands)	Vendor 1	4.5/10	1/5
	US English In-Car (commands, addresses, radio-controls, navigation, lifestyle information services and points-of-interest)	Vendor 1		1/5
	US English In-Car	Vendor 2		1/5
	German In-Car	Vendor 2		1/5
	Japanese In-Car	Vendor 2		1/5
Channel errors	1% BLER	Vendor 1	2/10	¼
	3% BLER	Vendor 1		¼
	1% BLER	Vendor 2		¼
	3% BLER	Vendor 2		¼

Table A2.2: Weighting of evaluation databases at 16kHz

Appendix 3: Illustration of recommendation based on relative improvement

	Relative improvement									
AMR error rate	**10.0**	**15.0**	**20.0**	**25.0**	**30.0**	**35.0**	**40.0**	**50.0**	**60.0**	**70.0**
40	36.0	34.0	32.0	30.0	28.0	26.0	24.0	20.0	16.0	12.0
35	31.5	29.8	28.0	26.3	24.5	22.8	21.0	17.5	14.0	10.5
30	27.0	25.5	24.0	22.5	21.0	19.5	18.0	15.0	12.0	9.0
25	22.5	21.3	20.0	18.8	17.5	16.3	15.0	12.5	10.0	7.5
20	18.0	17.0	16.0	15.0	14.0	13.0	12.0	10.0	8.0	6.0
18	16.2	15.3	14.4	13.5	12.6	11.7	10.8	9.0	7.2	5.4
16	14.4	13.6	12.8	12.0	11.2	10.4	9.6	8.0	6.4	4.8
14	12.6	11.9	11.2	10.5	9.8	9.1	8.4	7.0	5.6	4.2
12	10.8	10.2	9.6	9.0	8.4	7.8	7.2	6.0	4.8	3.6
10	9.0	8.5	8.0	7.5	7.0	6.5	6.0	5.0	4.0	3.0
9	8.1	7.7	7.2	6.8	6.3	5.9	5.4	4.5	3.6	2.7
8	7.2	6.8	6.4	6.0	5.6	5.2	4.8	4.0	3.2	2.4
7	6.3	6.0	5.6	5.3	4.9	4.6	4.2	3.5	2.8	2.1
6	5.4	5.1	4.8	4.5	4.2	3.9	3.6	3.0	2.4	1.8
5	4.5	4.3	4.0	3.8	3.5	3.3	3.0	2.5	2.0	1.5
4	3.6	3.4	3.2	3.0	2.8	2.6	2.4	2.0	1.6	1.2
3	2.7	2.6	2.4	2.3	2.1	2.0	1.8	1.5	1.2	0.9
2	1.8	1.7	1.6	1.5	1.4	1.3	1.2	1.0	0.8	0.6
1	0.9	0.85	0.80	0.75	0.70	0.65	0.6	0.50	0.40	0.30
0.5	0.5	0.43	0.40	0.38	0.35	0.33	0.3	0.25	0.20	0.15
0.1	0.1	0.09	0.08	0.08	0.07	0.07	0.1	0.05	0.04	0.03