Title: **Extended Mel-Cepstrum Front-End (XFE) and Extended Advanced Front-End (XAFE) Algorithms – Comparison of Test Results to Requirements**

Authors: **Alex Sorin (IBM), Tenkasi Ramabadran (Motorola)**

Date: **31-Mar-2003**

Version: **1.0**

Revision history:
Version 1.0: *Creation, March 31, 2003*

1. Scope

This document compares the different test results obtained for the Extended Mel-Cepstrum Front-End (XFE) and the Extended Advanced Front-End (XAFE) to their requirements.

2. Algorithm Documents

The XFE and XAFE algorithm descriptions are provided in the following documents:

a) *Au42203_XFE_Algorithm_Description*: Extended Mel-Cepstrum Front-End (XFE) Algorithm Description
b) *Au42303_XAFE_Algorithm_Description*: Extended Advanced Front-End (XAFE) Algorithm Description

3. Requirements Documents

The requirements of the XFE and XAFE algorithms are set forth in the following documents:

a) *Au41202_Extended_FE_Requirements_v1_4*: Requirements for front-end extension for tonal-language recognition and speech reconstruction
b) *Au41302_Reconstruction_Intelligibility_Testing_v1_2*: Intelligibility Testing of the Reconstructed speech synthesized using the DSR parameters
c) *Au41402_tonal_language_evaluation_v1_1*: Tonal-language recognition evaluation using the extended WI-030 front-end

4. Results Documents

The test results obtained for the XFE and XAFE algorithms are provided in the following documents:

a) *Au41803_Dyna_DRT_Report*: Intelligibility Testing – Diagnostic Rhyme Test (DRT) Report
b) *Au41903_Fan_TT_Report*: Intelligibility Testing – Transcription Test (TT) Report
c) *Au42003_TLR_Results_IBM*: Tonal-language recognition evaluation results obtained by IBM using the extended WI-030 front-end

d) *Au42103_TLR_Results_Motorola*: Tonal-language recognition evaluation results obtained by Motorola using the extended WI-030 front-end

5. Comparison of Results to Requirements

The different requirements of the XFE and XAFE algorithms are presented here in the same order they appear in the *Au41202* document and compared with the test results where appropriate.

5.1 Complexity

The overall complexity numbers for the extended front-ends are shown in table 1 below (reproduced from *Au41202*).

Table 1: Complexity requirements/objectives for the XFE and XAFE algorithms

Measure	Requirement
WMOPS	< 17
ROM size (kWords)	< 15
RAM size (kWords)	< 6

The numbers in table 1 represent the requirements for the XFE front-end. For the XAFE front-end, the numbers in table 1 represent the objectives (as was decided during August 2002 teleconference of the Aurora committee).

The actual complexity numbers for the XFE and XAFE front-ends are shown in tables 2 and 3 below.

Table 2: Complexity of the XFE algorithm

Measure	FE$^{\perp}$	Extension	XFE
WMOPS	6.28	5.62	11.90
ROM size (kWords)	1.88	5.65	7.53
RAM size (kWords)	1.92	3.81	5.73

$^{\perp}$ *Assumed to be one half the complexity of the Advance Front End*

Table 3: Complexity of the XAFE algorithm

Measure	AFE*	Extension	XAFE
WMOPS	12.55	5.21	17.76
ROM size (kWords)	3.752	5.396	9.148
RAM size (kWords)	3.830	3.290	6.864**

** From Motorola-France Telecom-Alcatel Advanced Front-End Proposal, January 31, 2002*
***Scratch memory reusing*

It is seen that the complexity numbers for the XFE algorithm are well below the permitted values whereas the complexity numbers for the XAFE algorithm are fairly close to the desirable values.

Thus, both the XFE and XAFE front-ends satisfy the complexity requirements.

5.2 Data Rate

The total data rate of the XFE and XAFE front-ends should not exceed 5800 bps for the different sampling rates, viz., 8, 11, and 16 kHz. This corresponds to an additional bit rate of 1000 bps for the extensions over the un-extended FE and AFE front-ends.

The actual data rate of the XFE and XAFE front-ends is 5600 bps for the different sampling rates corresponding to an additional bit rate of 800 bps for the extensions. The makeup of this additional bit rate is as follows: For every frame-pair, 16 bits (or 2 octets) are added – 12 bits for pitch information, 2 bits for class information, and 2 bits for CRC protection of pitch and class bits. Sixteen bits for every 20 ms corresponds to 800 bps.

Thus, both the XFE and XAFE front-ends satisfy the data rate requirement.

5.3 Update Rate

The update rate of the XFE and XAFE front-ends for transmission of the additional features, viz., pitch period and voicing class, should be the same as that of the standard cepstral features, i.e., 10 ms for all sampling rates.

The update rate of the XFE and XAFE front-ends for transmission of the additional features is 10 ms at all sampling rates.

Thus, both the XFE and XAFE algorithms satisfy the update rate requirement.

5.4 Bit-stream formatting and error protection

The requirements for "bit-stream formatting and error protection" as specified in *Au41202* are: The standard packaging should be extended to support transmission and error protection of the additional acoustic features. The multi-frame format (synchronization sequence and header field) should not change. Some of the 9 unused "expansion bits" EXP1-9 in the multi-frame header should be assigned to indicate whether additional acoustic information for reconstruction and tonal-language recognition is available in the stream. The CRC bits associated with a pair of frames may be extended (i.e., more than 4 CRC bits for a pair of frames).

In the XFE and XAFE algorithms, the standard packaging used by the FE and AFE algorithms has been retained. The multi-frame format along with the synchronization sequence and header field has been retained. The number of data (plus CRC) bits for each frame-pair has been extended from 92 bits to 108 bits, i.e., 16 extra bits (or 2 octets) have been added. The 16 bits are made up of the following: 12 bits for the pitch period information (7 bits for the first frame and 5 bits for the second frame), 2 bits for voicing class information (1 bit each for the two frames), and 2 CRC bits protecting the preceding 14 pitch and class bits. Thus, the 4 CRC bits used to protect the logE and cepstral features are left undisturbed. The EXP1 bit has been used to indicate that the front-end has been extended, i.e., a 0 bit indicates "no extension" and a 1 bit indicates "extension". Complete details regarding "bit-stream formatting and error protection" are found in the algorithm description documents *Au42203* and *Au42303*.

Thus, both the XFE and XAFE algorithms satisfy the "bit-stream formatting and error protection" requirements.

5.5 Terminal-side latency

The delay required for extraction of the additional features, viz., pitch period and voicing class, should not exceed the delay required for extraction of standard cepstral features. This means the look-ahead cannot be more than 60 samples at 8 kHz, 73 samples at 11 kHz, and 120 samples at 16 kHz sampling rates.

The XFE and XAFE front-ends do not introduce any additional delay beyond what is already present in the un-extended FE and AFE algorithms.

Thus, both the XFE and XAFE algorithms satisfy the terminal-side latency requirement.

5.6 Intelligibility of reconstructed speech

The requirements for the intelligibility of reconstructed speech are set forth in complete detail in the *Au41302* document. The intelligibility of the reconstructed speech synthesized by both XFE and XAFE algorithms are compared against the intelligibility of the speech synthesized by two well-known vocoders, viz., LPC10 vocoder and MELP vocoder both operating at 2400 bps. The performance of the LPC10 vocoder establishes the minimum requirement under nominal conditions, i.e., nominal signal level (-26 dBov), nominal sampling rate (8 kHz), and clean channel. The performance of the MELP vocoder establishes the objective ("nice-to-have" performance) under all tested conditions (different signal levels, sampling rates, and channel errors).

5.6.1 DRT

The primary intelligibility evaluation method is the Diagnostic Rhyme Test (DRT). The results of the many DRT's performed can be found in the *Au41803* document. In the *Au41803* document, the different conditions are identified by a Number-Letter-Number combination, e.g., 3A2. The first number corresponds to the Experiment (or table) number (I through IV) in Appendix II of the *Au41302* document. The letter corresponds to the row (A through E), and the second number corresponds to the column of the appropriate table in Appendix II of *Au41302*. So, the combination 3A2 refers to the condition in Experiment III, row A, column 2, which is XFE reconstruction under EP1 channel error condition. The summary DRT results from *Au41803* are reformatted here to correspond to the different experiments as specified in Appendix II of *Au41302*. The reformatted results are shown below in tables 4 – 7.

Table 4 (DRT Experiment I) – Background Noise Conditions

Noise Type: / Coder:	Clean	Car 10dB	Street 15dB	Babble 15dB
Unprocessed	95.7	95.5	92.4	93.8
XFE Reconstruction	93.0	88.8	85.0	87.1
XAFE Reconstruction	92.8	88.9	87.5	87.9
LPC-10	86.9	81.3	81.2	81.2
MELP	91.6	86.8	85.0	85.3

Table 5 (DRT Experiment II) – Input Signal Levels

Input level: / Coder:	-10dB	0 dB*	+10dB
XFE Reconstruction	91.7	93.0	93.3
XAFE Reconstruction	92.1	92.8	92.8

* From Experiment I

Table 6 (DRT Experiment III) – Channel Errors

Bit Errors: / Coder:	None*	EP1 (C/I 10 dB)	EP2 (C/I 7 dB)	EP3 (C/I 4 dB)
XFE Reconstruction	93.0	92.6	92.1	83.4
XAFE Reconstruction	92.8	92.6	92.0	83.4

* From Experiment I

Table 7 (DRT Experiment IV) – Sampling Frequencies

Sampling Frequency: / Coder:	8 kHz*	11 kHz	16 kHz
XFE Reconstruction	93.0	92.9	94.2
XAFE Reconstruction	92.8	93.5	92.1

* From Experiment I

In table 4, the DRT results for clean and background noise conditions are shown for the different coders. It is seen that the DRT scores of the XFE and XAFE reconstruction algorithms easily meet the minimum requirements set by the LPC10 coder. In fact, they are statistically equivalent to or better than the objectives set by the MELP coder.

In table 5, the DRT results for different signal levels are shown for the XFE and XAFE reconstruction algorithms. The objective here is that the DRT scores at +/- 10 dB signal levels are close to the DRT scores at nominal signal level. It is seen that the DRT scores increase slightly for +10 dB signal level, and decrease slightly for the –10 dB signal level. The results at +/- 10 dB, however, are quite close to the DRT scores at nominal signal level.

In table 6, the DRT results for different channel error conditions are shown for the XFE and XAFE reconstruction algorithms. The objective here is that the DRT performance degrade gracefully with increasing channel errors. It is seen that the performance is only slightly affected for C/I of 10 dB (EP1) and 7 dB (EP2) but to a large extent for C/I of 4 dB (EP3). This characteristic is similar to that of the recognition performance results of the un-extended FE and AFE front-ends.

In table 7, the DRT results for different sampling rates are shown for the XFE and XAFE reconstruction algorithms. The objective here is that the DRT scores at 11 kHz and 16 kHz are close to the DRT scores at 8 kHz. It is seen that this objective is satisfied for both XFE and XAFE algorithms.

Thus, both the XFE and XAFE algorithms meet or exceed the intelligibility requirements as evaluated by DRT.

5.6.2 TT

Another intelligibility evaluation method is the Transcription Test (TT). The results of the TT can be found in the *Au41903* document. In the *Au41903* document, the different combinations of message-transcriber identified by Mx-Tx (where x ranges from 1 through 5) are the same as shown in Table 7.1 of *Au41302*. The results from the *Au41903* document are compiled into tabular form in tables 8 and 9 shown below.

In table 8, the number of missed, wrongly transcribed, and partially transcribed words for the different coders under different background noise conditions are shown in rows 1 – 5, and columns 1 – 5. The number of words in each message (corresponding to each column) and the total number of words in all five messages (last column) are shown in the row below the table. The average percentage error rates (for missed, wrongly transcribed, and partially transcribed words) for all background noise conditions combined are shown in rows 1 – 5, last column. It is seen that the average word error rates for the XFE and XAFE reconstruction algorithms are lower than that of the LPC10 coder thus satisfying the requirement. In fact, the XFE and XAFE word error rates are lower than that of the MELP coder thus meeting the objective.

In table 9, the number of inserted words for different coders and background noise conditions are provided mainly for informational purposes.

Thus, both the XFE and XAFE algorithms meet or exceed the intelligibility requirements as evaluated by TT.

Table 8: Number of Missed, Wrongly transcribed, and Partially transcribed Words

Background noise: / Coder:	Clean	Car	Street	Babble	Clean	Average Error (%)
Uncoded (original)	1,1,2	1,0,1	0,2,4	3,9,3	0,4,1	0.5492
XFE reconstruction	1,6,1	0,3,6	2,9,4	5,9,2	1,4,5	0.9954
XAFE reconstruction	0,6,2	0,5,4	0,4,3	3,5,2	1,6,5	0.7894
LPC-10 coder	8,18,6	62,26,7	67,22,7	47,12,3	18,10,9	5.5260
MELP coder	0,3,1	1,6,3	4,6,2	16,10,3	1,9,5	1.2013
No. of words in message	1166	1153	1155	1149	1204	Total: 5827

Table 9: Number of Inserted Words

Background noise: / Coder:	Clean	Car	Street	Babble	Clean	Average Error (%)
Uncoded (original)	2	0	4	1	0	0.1201
XFE reconstruction	1	1	5	1	0	0.1373
XAFE reconstruction	2	2	1	1	1	0.1201
LPC-10 coder	4	3	1	2	3	0.2231
MELP coder	1	2	3	1	1	0.1373

No. of words in message	1166	1153	1155	1149	1204	Total: 5827

5.7 Tonal language recognition

The requirements for the tonal language recognition performance of the tonal features (pitch period and voicing class) generated by the XFE and XAFE algorithms are set forth in complete detail in the *Au41402* document. The objective here is to verify that the tonal features generated by the XFE and XAFE algorithms are comparable to the tonal features generated by proprietary algorithms in terms of tonal language recognition performance. The tonal language recognition evaluation has to be performed by at least two evaluators using MFCC based recognition engines and tone-sensitive databases both internal (proprietary) and common.

The results of the tonal language recognition evaluations performed by IBM and Motorola can be found respectively in the *Au42003* and *Au42103* documents. The overall results from Tables 1 and 2 (last column) of the *Au42003* document are shown in table 10 below. Similarly, the overall results from Tables 1, 2, 3, and 4 (last column) of the *Au42103* document are shown in table 11 below. In tables 10 and 11 below, the columns titled "Cantonese Digits" identify the common database. The remaining columns correspond to internal (or proprietary) databases.

Table 10: Tonal Language Recognition Evaluation Results from IBM

Configuration	Error Rate (%) OR Improvement (%)	
	Mandarin Digits	Cantonese Digits
C1 (no F0)	3.31	4.00
C2 (proprietary F0)	3.08	4.41
C3 (WI30 F0)	3.04	3.99
D_{21}	6.95	-10.25
D_{31}	8.16	0.25
D_{32}	1.30	9.52

Table 11: Tonal Language Recognition Evaluation Results from Motorola

Configuration	Error Rate (%) OR Improvement (%)			
	Mandarin Digits	Mandarin Commands	Cantonese Digits	
			Clean Training	Multi Training
C1 (no F0)	31.13	40.12	16.70	10.74
C2 (proprietary F0)	29.84	33.92	19.14	10.58
C3 (WI30 F0)	29.45	33.89	14.52	8.89
D_{21}	4.14	15.45	-14.61	1.49
D_{31}	5.40	15.53	13.05	17.23
D_{32}	1.31	0.09	24.14	15.97

In tables 10 and 11, it is seen that the relative improvements obtained from using the XFE tonal features over no tonal features (identified by the D_{31} row) are greater than the relative improvements obtained by using the proprietary tonal features over no tonal features (identified by the D_{21} row). In addition, the relative improvements obtained by using the XFE tonal features over the proprietary tonal features (identified by the D_{32} row) are positive. Therefore, the tonal features generated by the XFE algorithm are comparable in performance to the tonal features generated by the two proprietary algorithms evaluated. (Note: Since the XFE and XAFE algorithms for generating tonal features are practically identical, only the XFE features were evaluated).

Thus both the XFE and XAFE algorithms satisfy the requirements specified regarding tonal language recognition.

6. Conclusions

This document has compared the performance data contained in the "results documents" against the requirements for intelligibility and tonal language recognition performance specified in the "requirements documents". This document has also compared the algorithm description in the "algorithm documents" against the algorithm requirements for complexity, data rate, update rate, bit-stream formatting & error protection, and latency in the "requirements documents". The XFE and XAFE algorithms have been found to *meet or exceed* all the requirements set forth in the "requirements documents".