IEICE TRANS. INF. & SYST., VOL.E91–D, NO.6 JUNE 2008

LETTER

# Improved Frame Mode Selection for AMR-WB+ Based on Decision Tree

**Jong Kyu KIM**[†a)], *Nonmember and* **Nam Soo KIM**[†b)], *Member*

**SUMMARY** In this letter, we propose a coding mode selection method for the AMR-WB+ audio coder based on a decision tree. In order to reduce computation while maintaining good performance, decision tree classifier is adopted with the closed loop mode selection results as the target classification labels. The size of the decision tree is controlled by pruning, so the proposed method does not increase the memory requirement significantly. Through an evaluation test on a database covering both speech and music materials, the proposed method is found to achieve a much better mode selection accuracy compared with the open loop mode selection module in the AMR-WB+.

***key words:*** *AMR-WB+, audio coding, mode selection, decision tree*

## 1. Introduction

The AMR-WB+, which was standardized by 3GPP in 2004, is one of the state-of-art audio coding algorithms. In the standard selection phase, the AMR-WB+ showed a unique performance at low bitrates from below 10 kbps up to 24 kbps [1]. This qualified performance is largely due to the hybrid structure integrating both the speech and transform coding techniques. In hybrid audio coding, each audio frame is encoded by one of the multiple coding modes depending on the signal characteristics. In the AMR-WB+, there exist two coding modules, Algebraic Code Excited Linear Prediction (ACELP) and Transform Coded Excitation (TCX). Each coding module has been designed to fit to different types of audio signals. Therefore, the accurate selection for the coding mode is essential to achieve high coding efficiency.

Several methods have been proposed for coding mode selection in hybrid audio coding techniques. Closed loop methods select the coding mode by iteratively trying all possible coding modes such that the coding gain or other objective measures can be optimized [2]. In contrast, open loop methods make the decision based on features extracted from audio signal without actually synthesizing the signal [3]. Although closed loop mode selection is generally found to result in a better coding performance, high computational load is inevitable due to the repetitive encoding process, which makes it inappropriate for the small device applications where minimizing the computational cost is an important issue.

Manuscript received November 19, 2007.
Manuscript revised February 21, 2008.

[†]The authors are with School of Electrical Engineering and INMC, Seoul National University, Seoul, 151–744, Korea.
a) E-mail: ckkim@hi.snu.ac.kr
b) E-mail: nkim@snu.ac.kr

DOI: 10.1093/ietisy/e91–d.6.1830

In this letter, we propose an open loop mode selection algorithm based on decision tree. The decision tree is trained utilizing the result of the closed loop mode selection as a target classification such that the coding performance of the proposed method approaches that of the closed loop mode selection. In addition, the mode selection is designed to be carried out in a super-frame-wise manner instead of a frame-wise manner followed by post processing. In the proposed approach, the information in a super-frame is jointly exploited, thus avoiding additional computation invoked by the post process. Through an evaluation over various speech and audio data, it has been discovered that the proposed algorithm shows improved mode selection accuracy compared with the open loop mode selection algorithm implemented in the AMR-WB+ with acceptable memory requirement and comparable computational complexity.

## 2. Coding Mode Selection in AMR-WB+

In order to take advantage of both the speech and audio coding techniques, a number of hybrid audio coding structures have been investigated such as the Transform Predictive Coder (TPC) [4] and Transform Coded Excitation (TCX) [5]. The main purpose of hybrid audio coding is to improve coding efficiency in case of mixed contents containing both speech and music signals. For that reason, the target of mode selection is to accurately discriminate between speech and music signals [3]. Moreover, since abrupt mode change may cause audible artifacts, it is desirable to let the mode switching be made in a constrained manner [3], [6]. Recently, a hybrid audio coding technique was developed employing the window overlap scheme which combines the waveform coder and a subband coder without any constraint on mode switching [2]. It is also noted that a super-frame scheme based on larger window was proposed to enhance the coding gain of the wide-band audio signal by providing a variety of time-frequency resolution due to the increased number of coding modes. All of these techniques have been incorporated in the 3GPP AMR-WB+ standard [7].

In the AMR-WB+, each block of 1024 samples is specified as a super-frame. A single super-frame consists of four consecutive frames of 256 samples with additional overlapping region between adjacent frames. Each frame is encoded by one of four different coding modes: ACELP, TCX of 256, 512 and 1024 samples. With a few restrictions on the combination of coding modes, 26 different coding modes are allowed for each super-frame as shown in Table 1 where

0, 1, 2 and 3 denote the ACELP, TCX of 256, 512 and 1024 modes respectively.

In order to find the optimal coding mode, the standard encoder evaluates the coding efficiency for each coding mode in a closed loop manner. Since there are redundancies in mode combination, the closed loop mode selection is completed in 11 iterations as shown in Table 2 where each column represents a frame, each row indicates the iteration index and each entry denotes the encoding mode to be evaluated. The coding efficiency is evaluated in terms of the perceptually weighted signal to noise ratio (SNR).

In addition to the closed loop mode selection, an open loop mode selection module is also provided as a low complexity alternative. In the open loop mode selection, coding mode is selected for each frame between ACELP and TCX of 256 samples using features such as the voice activity decision (VAD) flag, ratio of low frequency energy to high frequency energy, spectral difference between adjacent frames, long term prediction gain and so on. Following the preliminary discrimination between ACELP and TCX, a post process to select the coding mode of the whole super-frame is conditionally invoked to decide among the TCX modes with various window lengths, which is referred to as TCX Selection (TCXS). TCXS is also conducted in the same manner as in the closed loop mode selection.

In the performance characterization of 3GPP audio codecs [8], the encoder complexity of the open loop mode selection is reported to be around half of that of the closed loop mode selection. However, the subjective audio quality of the open loop mode selection has been found to be significantly worse compared with the closed loop mode selection. These results naturally lead to an effort to develop an open loop mode selection method with better audio quality.

**Table 1** Coding modes in the AMR-WB+ super-frame.

| | | | |
|---|---|---|---|
| (0, 0, 0, 0) | (0, 0, 0, 1) | (2, 2, 0, 0) | |
| (1, 0, 0, 0) | (1, 0, 0, 1) | (2, 2, 1, 0) | |
| (0, 1, 0, 0) | (0, 1, 0, 1) | (2, 2, 0, 1) | |
| (1, 1, 0, 0) | (1, 1, 0, 1) | (2, 2, 1, 1) | |
| (0, 0, 1, 0) | (0, 0, 1, 1) | (0, 0, 2, 2) | |
| (1, 0, 1, 0) | (1, 0, 1, 1) | (1, 0, 2, 2) | |
| (0, 1, 1, 0) | (0, 1, 1, 1) | (0, 1, 2, 2) | (2, 2, 2, 2) |
| (1, 1, 1, 0) | (1, 1, 1, 1) | (1, 1, 2, 2) | (3, 3, 3, 3) |

**Table 2** Coding modes in a super-frame of the AMR-WB+.

| | Frame0 | Frame1 | Frame2 | Frame3 |
|---|---|---|---|---|
| 1 | ACELP | | | |
| 2 | TCX256 | | | |
| 3 | | ACELP | | |
| 4 | | TCX256 | | |
| 5 | TCX512 | TCX512 | | |
| 6 | | | ACELP | |
| 7 | | | TCX256 | |
| 8 | | | | ACELP |
| 9 | | | | TCX256 |
| 10 | | | TCX512 | TCX512 |
| 11 | TCX1024 | TCX1024 | TCX1024 | TCX1024 |

## 3. Mode Selection Method Based on Decision Tree Training

As a general classifier for open loop mode selection, we apply the decision tree built over a training data. Each training sample contains a feature vector and the corresponding closed loop mode selection result as the target label.

Decision tree classifies an input feature vector by passing it through test nodes from the root down to a certain leaf node marked with a classification label as shown in Fig. 1. At each test node, a decision is made to assign the input feature vector to one of its child nodes. As a consequence, the leaf node occupies a non-overlapping region in the feature space as shown in Fig. 2 where training samples, decision boundaries and mode selection results are illustrated.

In general, the accuracy of a classifier evaluated on the training data increases with the number of relevant features. However, since the increased number of features requires more computation, it is important to exploit maximum amount of information from the features. In the proposed method, the features used in the open loop mode selection of the AMR-WB+ are employed to construct an 81st-order super-frame feature vector: four sets of 20 features from each frame and the maximum energy among the four frames. By integrating the features from all the frames

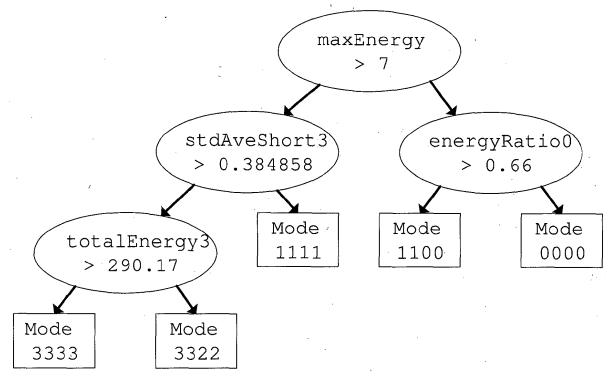


**Fig. 1** An example of decision tree classifier.

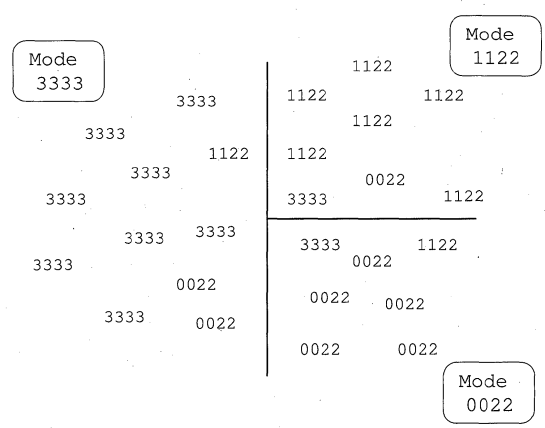


**Fig. 2** An example of feature space divided by tests.

that form a super-frame mode selection can be done in a single step instead of the two stage framework implemented in the AMR-WB+.

In this letter, decision tree is constructed using the Iterative Dichotomiser 3 (ID3) algorithm which is known to generate a smaller and more accurate classifier [9]. ID3 grows a decision tree by recursively splitting each node based on a set of decision rules. For a nominal feature, the test made at each node is a selection among several possible values while for numerical feature, the test is to check whether the feature is greater or less than a constant. In order to measure the effectiveness of a test, the gain ratio is adopted as a splitting criterion. If $D$ indicates a set of feature vectors assigned to a node, and $C$ is the number of classes representing selection modes, then the uncertainty of the distribution of feature vectors is measured in terms of entropy given by

$$Info(D) = -\sum_{j=1}^{C} p(D, j) \times \log_2(p(D, j)) \tag{1}$$

where $p(D, j)$ denotes the proportion of feature vectors in $D$ which belong to the $j_{th}$ class. If $D$ is split into $k$ child nodes according to a test, $T$, the information gain is computed as follows:

$$Gain(D, T) = Info(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times Info(D_i) \tag{2}$$

where $D_i$ is the set of feature vectors assigned to the $i_{th}$ child node and $|\cdot|$ represents the cardinality of a set. The information gain implies the decreased uncertainty caused by the splitting. Therefore, the larger the informational gain is, the more efficient the test is considered to be. However, since the information gain is strongly affected by the number of split nodes, it should be penalized by the split information given by

$$Split(D, T) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} \times \log_2 \frac{|D_i|}{|D|}. \tag{3}$$

Among the possible tests, the one with maximum gain ratio is chosen as the test of the node and the training samples are split into child nodes according to it.

To make the induced decision tree feasible for practical applications, we usually apply pruning. The size of a decision tree rapidly grows with the order of feature vector. As will be shown in Section 4, the induced tree with 81st-order feature vectors is excessively large. Although the computation for run-time classification is almost negligible in the decision tree, the large memory requirement might be an obstacle for small device applications. Hence, after the decision tree is induced, subtree replacement is applied which shrinks the size of the tree.

In subtree replacement, some subtrees are replaced by single leaves. To decide whether to carry out the replacement in a subtree, the error rate $q$ is estimated at each node over the training samples in the subtree. Let $E$ denote the number of erroneous samples out of $N$ training samples in a subtree. Then, $e = E/N$ is the observed error rate and the confidence interval for $e$ with a confidence level $c$ is defined by

$$P\left[-z \le \frac{e - q}{\sqrt{q(1-q)/N}} \le z\right] = c \tag{4}$$

based on the assumption that $e$ follows a normal distribution. With (4), the estimated error rate, $\hat{q}$ is obtained using the upper confidence bound $z$ as follows:

$$\hat{q} = \frac{e + \frac{z^2}{2N} + z\sqrt{\frac{e}{N} - \frac{e^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}. \tag{5}$$

The error rate $q$ is estimated for each leaf node in a subtree and weighted-summed according to the number of samples in each node. Then, it is compared with $\hat{q}$ obtained in the parent node. Subtree replacement is taken place when the estimated error rate of the parent node is lower.

## 4. Experimental Results

In order to implement the mode selection method based on decision tree, training set was obtained from audio materials ranging from speech to various genres of music. The length of the training set was 1 hour containing 30 minutes of speech and 30 minutes of music materials. In order to make the trained decision tree unbiased, we incorporated speech and music material equally to the sample set. The specification of the training set is described in Table 3.

The accuracy of mode selection was evaluated through a cross validation strategy. In $n$-fold cross validation, the whole training data is randomly divided into $n$ subsets of almost equal size. At each evaluation trial, $n - 1$ subsets are used to train the decision tree and the remaining subset is applied to evaluating the performance. The training and evaluation phases are carried out $n$ times for each separate subset and the test result is averaged over the $n$ subsets. The result from 9-fold cross-validation test is shown in Table 4 where the accuracy of mode selection is compared to that of the open loop mode selection in the AMR-WB+. The accuracies of the super-frame mode and the individual frame are separately given. The accuracy of the proposed method was found to improve the classification rate by 9.78% and 6.39% for super-frame and individual frame, respectively.

For the purpose of investigating the practical feasibility of this method, the number of nodes in the induced trees are given in Table 5. As shown in the table, the number of tree nodes decreased from 9933 to 949 by the pruning process. If $N_n$ and $N_l$ respectively denote the number of nodes

**Table 3** Test specifications.

| | |
|---|---|
| Number of Channels | Stereo |
| Input Audio Sampling Rate | 25,600 Hz |
| Encoding Bitrate | 16 kbps |
| Number of Train Instances | 45000 |
| Length of Train Instances | 1 hour |

**Table 4** Comparison of mode selection accuracies.

| Method | Super-frame | Frame |
|---|---|---|
| AMR-WB+ Open Loop Method | 60.16% | 73.41% |
| Proposed Method | 69.94% | 79.80% |

**Table 5** The size of trees and mode selection accuracy.

| | With Pruning | Without Pruning |
|---|---|---|
| Number of Nodes | 949 | 9933 |
| Number of Leaves | 524 | 4967 |
| Mode Selection Accuracies | 69.94% | 66.02% |

and leaves, the table size for storing the decision tree becomes $N_n(w_f + 2w_p) + N_l w_m$ where $w_f$, $w_p$ and $w_m$ denotes the number of words required to store a constant, pointers to child nodes in a node, and the selected mode in a leaf node. Hence, in the experiment the number of words to store the decision tree was 4320 when a constant was expressed in 2 words and the pointers in 1 word. Considering that the size of the fixed table is 16142 words in the AMR-WB+, the reduced size of the tree does not significantly increase the memory requirement and is acceptable to the applications which have limited memory size [10]. In addition, the increased accuracy of the pruned tree indicates the improved generalization capacity.

## 5. Conclusion

In this letter, we have proposed a coding mode selection method for the AMR-WB+ audio coder based on decision tree. The decision tree classifier is trained to pursue the coding efficiency of the closed loop mode selection while maintaining the complexity of the open loop mode selection. Through a test and evaluation on either speech or music materials, the proposed method has been found to achieve a better accuracy than the open loop mode selection in the AMR-WB+. In addition, the size of the decision tree is kept sufficiently small such that the memory requirement is acceptable.

## Acknowledgments

This work was supported in part by Seoul R&BD Program (10544) and the Korea Science and Engineering Foundation (KOSEF) grant funded by Korea government (MOST) (No. R0A-2007-000-10022-0).

### References

[1] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: A new audio coding standard for 3rd generation mobile audio services," Proc. ICASSP 2005, vol.2, pp.1109–1112, March 2005.

[2] B. Bessette, R. Lefebvre, and R. Salami, "Universal speech/audio coding using hybrid ACELP/TCX techniques," Proc. ICASSP 2005, vol.3, pp.301–304, March 2005.

[3] L. Tancerel, S. Ragot, V.T. Ruoppila, and R. Lefebvre, "Combined speech and audio coding by discrimination," Proc. IEEE Workshop on Speech Coding 2000, pp.154–156, Sept. 2000.

[4] J. Chen and D. Wang, "Transform predictive coding of wideband speech signals," Proc. ICASSP 1996, vol.1, pp.275–278, May 1996.

[5] R. Lefebvre, R. Salami, C. Laflamme, and J-P. Adoul, "High quality coding of wideband audio signals using Transform Coded eXcitation (TCX)," Proc. ICASSP 1994, vol.1, pp.193–196, April 1994.

[6] B. Bessette, R. Salami, C. Laflamme, and R. Lefebvre, "A wideband speech and audio codec at 16/24/32 kbit/s using hybridACELP/TCX techniques," Proc. IEEE Workshop on Speech Coding 1999, pp.7–9, June 1999.

[7] 3GPP TS 26.290, "Audio codec processing functions; Extended adaptive multi-rate - Wideband (AMR-WB+) codec; Transcoding functions," Release 7, 2007.

[8] 3GPP TS 26.234, "Performance characterization of 3GPP audio codecs," Release 7, 2007.

[9] J.R. Quinlan. "Improved use of continuous attributes in c4.5," Journal of Artificial Intelligence Research, vol.4, pp.77–90, 1996.

[10] 3GPP TS 26.273, "Extended adaptive multi-rate - Wideband (AMR-WB+) codec; Fixed-point ANSI-C code," Release 7, 2007.