# A FUZZY APPROACH TOWARDS PERCEPTUAL CLASSIFICATION AND SEGMENTATION OF MP3/AAC AUDIO

*Serkan Kiranyaz, Ahmad Farooq Qureshi, Moncef Gabbouj*

Institute of Signal Processing, Tampere University of Technology, Tampere, Finland
{serkan, qureshi}@cs.tut.fi, moncef.gabbouj@tut.fi

## ABSTRACT

The paper presents a novel perceptual based fuzzy approach towards classification and segmentation for MP3 and AAC audio in the compressed domain. The input audio is split into segments, which are classified as speech, music, fuzzy or silent. The proposed method minimizes critical errors of misclassification by fuzzy region modeling, thus increasing the efficiency of both pure and fuzzy classification. The experimental results show that the critical errors are minimized and the method is robust to capturing and encoding parameters of MP3 and AAC bit streams. Due to the efficiency obtained from fuzzy-region modeling and improved accuracy via rule-based semantic approach, the method is designed specifically for the audio-based multimedia indexing and retrieval systems.

## 1. INTRODUCTION

Audio information has been recently used for content-based multimedia indexing and retrieval systems. As for audio segmentation and classification, several methods have been recently reported. Although audio classification has been mostly realized in uncompressed domain but with the emerging MPEG audio content, several methods have been reported for audio classification on MPEG-1 (Layer 2) encoded audio bit-stream. The last years have shown a widespread usage of MPEG Layer 3 (MP3) audio [3], [4] as well as proliferation of several video content carrying MP3 audio. The ongoing research on perceptual audio coding yield to a more efficient successor called (MPEG-2/4) Advanced Audio Coding (AAC) [4]. AAC has various similarities with its predecessor but promises significant improvement in coding efficiency.

This paper describes a fuzzy approach towards perceptual audio classification and segmentation directly from MP3 and AAC bit-streams. The process is a self-adapted technique, which logically builds on the extracted information throughout its execution, to produce a reliable result at the end. The proposed method proceeds in logical hierarchic steps and iterations, based on certain perceptual rulings that are applied on the basis of perceptual evaluation of the classification features and the behavior of the process. A typical example is, applying a threshold to an extracted feature-value for classification.

The proposed scheme is specifically designed for audio-based multimedia indexing and retrieval systems and currently used within MUVIS system [2]. The proposed method is unsupervised which does not get any feedback from video and therefore can be applied to any standalone MP3/AAC audio clip or to any media primitive that carries MP3/AAC audio. The method is also designed to provide global and reliable solutions for the various capturing/encoding parameters and modes such as sampling frequencies (i.e. 8KHz up to 48 KHz), channel modes (i.e. mono, stereo, etc.), compression bit-rates (i.e. 8kbps up to 448kbps), sound volume level, file types, etc.

For each audio segment the classification results into *speech, music, fuzzy* or *silent*. *Speech* and *music* are the pure class types. Class type of a segment can be defined as *fuzzy* if either it is not classifiable as a pure class type due to some significant background noise or it may be mixed by more than one pure class type. In MUVIS system [2], the audio information is indexed according to their classification types and during the retrieval process a crucial improvement can be achieved by comparing only the matching class types with each other (i.e. only speech with speech segments). Therefore, this high-level content analysis integrated into the indexed scheme will ensure to minimize the erroneous aural query retrievals with mismatched class types. For instance an audio segment with pure class content is only searched throughout the associated segments of the audio items in the database having the same (matching) pure class type such as *speech* or *music*. However, *fuzzy* content is to be compared with all the contents of the database (i.e. *speech*, *music* and *fuzzy)* since it might, by definition, contain various class types, background noise, aural effects, etc. Therefore, for the proposed method, any erroneous classification on pure classes is intended to be detected as *fuzzy,* so as to avoid significant retrieval errors (mismatches) due to the misclassification. In this context, three prioritized error types of classification, illustrated in Figure 1, are defined:

- **Critical Errors:** These errors occur when one pure class is misclassified into another pure class. Such errors significantly degrade the overall performance of an indexing/retrieval scheme.
- **Semi-critical Errors:** These errors occur when a *fuzzy* class is misclassified into a pure class. These errors moderately affect the performance of retrieval.
- **Non-critical Errors:** These errors occur when a pure class is misclassified as a fuzzy class. The effect of such errors on the overall indexing/retrieval scheme is negligible.

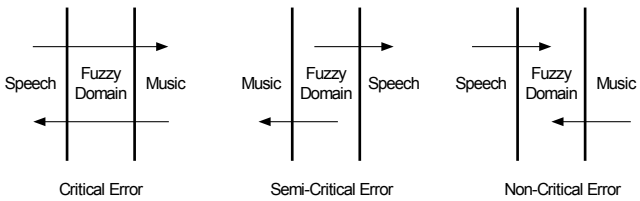


**Figure 1: Different error types in classification.**

## 2. FORMALIZATION OF COMPRESSED DOMAIN AUDIO FEATURES

The method uses the compressed domain audio features in order to perform classification and segmentation directly from the compressed bit-stream. The compressed domain features are *Total Frame Energy (TFE)*, *Band Energy Ratio (BER)*, Fundamental Frequency (FF) and Subband Centroid (SC) frequency. The formalization of such audio features is based on the formation of a generic MDCT sub-band template. Once the MDCT template formation is completed then the proposed algorithm can be applied to both types of bit-streams independent from the underlying encoding scheme. The details on the formation of MDCT template from MP3/AAC bit-stream and feature extraction in compressed domain can be found in [1].

## 3. MP3/AAC CLASSIFICATION AND SEGMENTATION

Audio segmentation and classification are mutually dependent problems. A good segmentation demands good classification and vice versa. Therefore, without any prior knowledge or supervising mechanism, the proposed algorithm proceeds in an iterative way, starting from granule/frame based classification and initial segmentation, to ensure a global segmentation and thus a successful classification per segment at the end. Figure 2 illustrates our 4-steps iterative approach to the audio classification and segmentation problem.

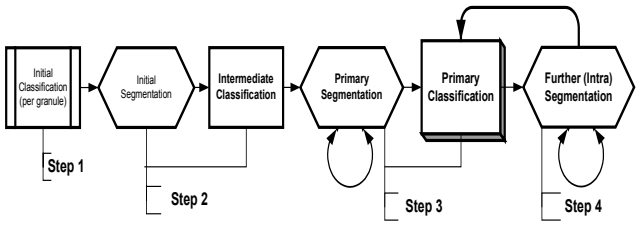


**Figure 2: The flowchart of the proposed method**

### 3.1. Initial Classification (per granule/frame)

In this step, each granule/frame is classified as *speech*, *music* or *silent*. Firstly, silence detection is performed on the basis of *Total Frame Energy* (*TFE*) [1] in two versions; with and without the pre-filtering of frequencies less than 80 Hz for each subband within every granule. This is done to avoid the low frequency microphone buzzing noise and other artifacts present in the ordinary recordings. Perceptually, the pre-filtering improves the overall content for a better analysis and also improves the performance of a feature used for classification (i.e. Pause Rate) by adding more silent granules (that were previously used to accommodate noise) within a segment to improve the speech classification. On the other hand, it does not degrade the quality of the content as normal human speech fundamental frequencies lie above 80 Hz [5].

If the granule/frame is not classified as silent, the *Band Energy Ratio (BER)* [1] is then calculated for a frequency of 500 Hz (most of speech energy is concentrated below 500Hz). If *BER* is over a threshold (i.e. 2%) the granule/frame is classified as *music*, otherwise *speech.*

### 3.2. Segmentation and feature extraction (per segment)

In this step, silent and non-silent segmentations are performed. As all the silent granules/frames have already been found in the previous step, the silent granules/frames are merged to form silent segments if enough number of silent granules/frames merge together to form a segment of larger duration than a preset threshold value (i.e. *0.2s*ec). All parts left between silent segments are then considered as non-silent segments. Once all non-silent segments are formed, then the classification of these segments is performed using the following features:

- ***Dominant Band Energy Ratio (DBER):*** For each non-silent segment, the dominant classifier type; the greater number of granule/frame types based on *BER* classification (in previous step), will determine the segment type.
- ***Pause Rate (PR):*** *PR* is the ratio between the number of *silent* granules/frames to the total number of granules/frames in a non-silent segment. The pre-filtered granules (as explained above) are used for this ratio. If *PR* is over a threshold ($T_{PR} = 2\%$), the segment is classified as a s*peech* segment, otherwise *music* by PR in this step.

### 3.3. Merging and Global Classification

After step 2 (section 3.2) is performed, small silent segments are eliminated and neighbor non-silent segments are merged where details about this process are given in [1]. The remaining non-silent segments are global enough to contain a single class type, which is further important in order to run other statistical features based on *Fundamental Frequency (FF)* estimation and *Sub-band Centroid (SC)* frequency extraction and *Pause Rate (PR)*. *FF* estimation [1] is performed per granule/frame. If no fundamental frequency is found over a granule/frame, typically that granule/frame is assumed to contain non-harmonic (unvoiced) components. *SC* frequency extraction [1] is the detection of the first moment of the spectral distribution (spectrum) or in compressed domain it can be estimated as the balancing frequency value for the absolute MDCT values. It is performed for all non-silent segments while excluding the silent granules/frames.

The *FF* feature used for classification of a segment is the product of mean *FF* with the percentage of harmonic granules/frames within a segment. The *SC* feature used to perform classification is the standard deviation of *SC* in a segment. In some cases, the mean of *SC* within a segment is also used for complementing standard deviation for classification. These features are extracted by smoothly sliding a short window through the entire length of the non-silent segment. The standard deviation of the *SC* is calculated using local windowed mean and global variance of *SC* in the segment. The third statistical feature used for segment classification is *PR*.

All possible feature-values of a feature make the feature-domain. This feature-domain can be divided (by the application of various thresholds) into parts for different class types, i.e. the

classification of a segment depends upon the part of the feature-domain in which the feature-value falls. This partitioning and thresholding process is purely based on aural perception and is implemented on the *PR*, *FF* and *SC* feature-domains. Within a segment, if *PR* is less than a threshold (i.e. 2%) the segment is classified as music, otherwise speech. Similarly, if the *FF* feature is lower than a threshold (i.e. 5000) the segment is classified as *speech*, otherwise *music*. Classification based on *SC* feature is *speech* if the feature-value is above a high (speech) threshold and *music* if below a low (music-) threshold. As mentioned before the feature domain between speech and music threshold is called fuzzy region. The class type of a segment is *fuzzy* if that particular feature value is within this *fuzzy* region.

The aforementioned features are also capable of forced-classification. The forced-classification is applicable on a segment only when an individual classification via a particular feature is utmost reliable in perceptual sense. For example if *PR* within a segment is above a high threshold value (i.e. 8%), the segment will have a forced-class type of *speech*. Similarly, if the result given by *FF* feature is over a maximum *FF* threshold, the segment will have a forced-class type of music. Likewise, if the overall segment *SC* standard deviation is above a maximum threshold (while *SC* mean is less than a min threshold), it will have a forced-class type of *speech* and it will have forced-class type of *music* if the *SC* standard deviation is less than an intermediate music threshold (while *SC* mean is greater than a threshold) or *SC* standard deviation is less than a minimum music threshold.

The final classification decision over a segment depends on its length. For a very small segment (less than 2sec) this decision is music if *FF* forced-classification exists for that segment. This is due to the high reliability of this classification even in smaller segments. In all other cases, the final decision is speech because of the fact that it is highly unlikely for music segment to be less than 2sec. For a long enough segment (> 2sec) the classification results based on each feature (*PR, FF and SC*) are fed into a three-step decision making process as explained next. This decision making process is entirely made up of the perceptual rulings which are logically arranged depending upon the reliability and behavior of every feature.

#### *3.3.1. Initial Decision*

The initial classification results are based on the forced-classification of a segment, only if forced classification exists for any of the three classification features. To avoid a final misclassification result in case two or more features force different class types for the same segment, a forced-decision-model is applied. Whenever *FF* feature applies a forced-music classification, it has the highest priority since the forced classification by the *FF* feature is the most reliable. Otherwise, whichever classifying feature, in its feature-domain, is relatively higher into the region of forced-classification with respect to its threshold value (more in the surety region) is chosen to be the decider.

If forced classification exists for a segment, the rest of the decision making process is simply bypassed (overruled), keeping the "forced" decision as the final classification. Otherwise, the intermediate and final decision steps are carried out on this segment.

#### *3.3.2. Intermediate Decision*

The intermediate decision is based on *PR* classification and *FF* classification. As *PR* feature is unreliable in the discrimination between *music* and *speech* with background *music* or noise, and similarly *FF* feature is unreliable in the discrimination between *speech* and *fuzzy* (*speech* with background music or noise), both of the features only classify a segment into speech or music class types. If both agree upon the classification type of a segment as *speech* or *music*, it is awarded that class type. In all other cases the intermediate decision of a segment is kept as *fuzzy*.

#### *3.3.3. Final Decision*

The final decision is composed of initial decision if a forced-classification exists for that segment. Otherwise, this decision is based on the intermediate decision and *SC* classification. For the final decision making process, a fuzzy region modeling is introduced, which is capable of validating and overruling the previous classification decisions to improve the results.

The intermediate decision is not changed if *SC* classification is a pure class. Otherwise, if *SC* classification is *fuzzy*, the final decision depends both upon the intermediate decision, *fuzzy region modeling* and the position of the *SC* standard deviation feature-value in this *fuzzy* region of the *SC* feature domain.

If the intermediate decision is a pure class, the final decision is *fuzzy* only for the segments with *SC* standard deviation feature-value above 15% (for music) or below 95% (for speech) of the fuzzy region length from the lower fuzzy region threshold. This is the *contracted fuzzy region model,* which ensures that only the segments with *SC* standard deviation feature-value closer to the center (region of surety) of the fuzzy region are re-classified as *fuzzy* and otherwise intermediate decision is not changed.

Similarly, if both intermediate decision and *SC* classification agree to be *fuzzy*, the intermediate decision is changed to *speech* or *music* depending upon the *expanded fuzzy region model*, i.e. above 99% (for speech) or below 3% (for *music)* of the fuzzy region length from the lower fuzzy region threshold. This ensures that classification of only those segments is overruled whose *SC* standard deviation feature-value is closer to the pure class boundaries (region of surety) of the fuzzy region.

### 3.4. Sub-Segment Analysis

Once final classification and segmentation is finished in step 3 (section 3.3), non-silent segments consisting of two or more sub-segments (without any silent part in between) might still need to be portioned into new segments. Therefore, a further segmentation is performed in order to separate sub-segments, which are not separated by silent parts.

The first part in this step tests if the length of the non-silent segment is significantly long (larger than a threshold). In this case *inner-breakpoints detection* is used to detect the new breakpoints within such segments, which is the real limit between two different sub-segments without a silent part between them. Within a sufficiently long segment, this algorithm firstly detects all the *music* granules/frames that have windowed-*SC* standard deviation value less than a threshold.

Beginning from the start of the segment, this algorithm locates first music granule/frame. Traversing from this granule/frame it tries to form a music sub-segment limit by reaching a non-music granule/frame or the end of the segment. In this case, if the sub-segment's length is less than the noise-length threshold, the limit granule/frame is not considered, and the algorithm continues looking for proper sub-segment boundaries. In other case, if this sub-segment's length is larger than noise-threshold, it is considered a sub-segment and *roll-down algorithm* is applied. The *roll-down algorithm* finds the first lowest point on the windowed-*SC* standard deviation curve within a short window, around the detected breakpoint, within the segment. This lowest point is the real limit between *speech* and *music* without a silent part between them. The breakpoint is validated only if the resulting segment's length is larger than a given music-length threshold. Otherwise, the large segment is kept unchanged and the algorithm stops.

After the inner breakpoints are saved, a *re-segmentation* method creates a new segment out of the new breakpoints, only if the classification of the new segment (based on step 3.3) is different from the previous. Otherwise the class type of this segment is kept unchanged.

After the re-segmentation, there might still be some small non-silent segments that are suspected to be misclassified if their classification differs from their nearby neighbors. For such small segments having at least one close enough neighbour (distant less than a threshold), a so-called *collateral decision* is applied. If this is the case, the classification of the short segment is changed depending upon the class of the neighbor segment.

Finally a merging operation is performed on all small silent segments (the duration is less than a threshold) having matching neighboring segments.

## 5. SIMULATION AND RESULTS

Experiments are carried out on standalone *MP3, AAC* audio clips, AVI and MP4 files containing MPEG-4 video along with *MP3* or *AAC* audio. These files contain diverse contents from TV channels showing News, Cartoon, Talk Show, Music Clips and Commercials or ordinary MP3 clips downloaded from Internet. The duration of clips are varying between 1-5 minutes up to 2 hours. The clips are recorded using several sampling frequencies from 16 KHz to 44.1 KHz so that both MPEG 1 and MPEG 2 phases are tested for Layer 3 audio. Both MPEG-4 and MPEG-2 AAC are recorded with the *Main* and *Low Complexity* profiles (object types). TNS (Temporal Noise Shaping) and M/S coding schemes are disabled for AAC. Around 70% of the clips are stereo and the rest is mono.

In total measures, the method is applied onto 120 (~8 hours) MP3 and 50 (~4 hours) AAC clips. Table 1 presents the error distribution for the proposed method separately on MP3 and AAC clips while Table 2 shows the segmentation results.

**Table 1: Error Distributions**

| | *Speech* | | *Music* | | *Fuzzy* |
|---|---|---|---|---|---|
| | **Critical** | **Non-Critical** | **Critical** | **Non-Critical** | **Semi-Critical** |
| **MP3** | 2.5% | 11.6% | 5.4% | 64% | 17.5% |
| **AAC** | 0.1% | 13.3% | 5.7% | 16.9% | 15.3% |

## 6. CONCLUSIONS AND FUTURE WORK

As given in Table 1, the simulation results show that both for MP3 and AAC, critical and semi-critical errors are minimized. Further, as intended, most of the occurred errors are bundled into non-critical error. The results prove that the perceptual approach adopted for the proposed algorithm is an appropriate answer to such a problem of classification and segmentation. Furthermore, as the classification features in the compressed domain are not ideal, the hierarchical steps and iterations through which the proposed algorithm runs, is probably the best way to proceed for a method to produce reliable results, if and when run in a generic form on MP3 or AAC files carrying any kind of audio content.

**Table 2: Segmentation Accuracy Results**

| *Segmentation Accuracy* | | |
|---|---|---|
| | **Normal Segmentation** | **Intra-Segmentation** |
| **MP3** | 98.25% | 76% |
| **AAC** | 97.9% | 79% |

Table 2 shows the segmentation results for MP3 and AAC audio. The efficiency of normal segmentation (segmentation based on silent granules/frames) and intra-segmentations (further segmentation of a segment into sub-segments) are encouraging although there is still possibility of improvement.

There is an ongoing study to further optimize this segmentation and classification method so as to further reduce the errors, improve the parameters that directly affect the overall results of the algorithm and thus further enhance its efficiency in all respects. Once these goals are met, the proposed method may be used as an advance stand-alone MP3/AAC audio analysis tool.

## 5. REFERENCES


[1] S. Kiranyaz, M. Aubazac, M. Gabbouj, "Unsupervised Segmentation and Classification over MP3 and AAC Audio Bit-streams", *WIAMIS Workshop*, pp. 338-345, London, 2003.

[2] S. Kiranyaz, K. Caglar, O. Guldogan, and E. Karaoglu, "MUVIS: A Multimedia Browsing, Indexing and Retrieval Framework", *Proc. Third International Workshop on Content Based Multimedia Indexing, CBMI 2003,* Rennes, France, 22-24 September 2003.

[3] D. Pan, "A tutorial on MPEG/Audio Compression", *IEEE Multimedia*, pp 60-74, 1995.

[4] Karl-Heinz Brandenburg, "MP3 and AAC Explained", *AES 17th International Conference*, Florence, Italy, September 1999.

[5] Baken, R. J. (1987). Clinical Measurement of Speech and Voice. London: Taylor and Francis Ltd.