A CONSTRUCTION OF COMPACT MFCC-TYPE FEATURES USING SHORT-TIME STATISTICS FOR APPLICATIONS IN AUDIO SEGMENTATION

Dirk von Zeddelmann and Frank Kurth

Research Establishment for Applied Science (FGAN)
Research Institute for Communication, Information Processing and Ergonomics (FKIE)
Neuenahrer Str. 20, 53343 Wachtberg, Germany
phone: + (49) 228-9435868, fax: + (49) 228-856277, email: {zeddelmann,kurth}@fgan.de

ABSTRACT

In this paper, we propose a new class of audio feature that is derived from the well-known *mel frequency cepstral coefficients* (MFCCs) which are widely used in speech processing. More precisely, we calculate suitable short-time statistics during the MFCC computation to obtain smoothed features with a temporal resolution that may be adjusted depending on the application. The approach was motivated by the task of audio segmentation where the classical MFCCs, having a fine temporal resolution, may result in a high amount of fluctuations and, consequently, an unstable segmentation. As a main contribution, our proposed *MFCC-ENS* (MFCC-Energy Normalized Statistics) features may be adapted to have a lower, and more suitable, temporal resolution while summarizing the essential information contained in the MFCCs. Our experiments on the segmentation of radio programmes demonstrate the benefits of the newly proposed features.

1. INTRODUCTION

The choice of suitable audio features is crucial for the most tasks in the field of audio information retrieval. Considering the task of *audio segmentation*, where a target signal is to be partitioned into a sequence of temporal segments, each being assigned a label such as *Speech* or *Music*, the *temporal resolution* of the underlying features is of particular importance.

To motivate the subsequently proposed new class of temporally adaptive features, we consider the particular problem of segmenting an audio signal recorded from a radio broadcast into the classes of *Music (C1)*, *Speech (C2)* and *Speech+Music (C3)*. A fourth class will be implicitly assumed for temporal segments which are not assigned any of the other class labels during the segmentation process.

As an example, Fig. 1 shows an excerpt of a radio programme consisting of three subsequent segments of speech, music and speech again. A correct segmentation hence would be a sequence of three segments labeled *C2*, *C1* and *C2*. The spectrogram (a) shows a time-frequency representation of the audio signal, the extracted MFCC-features are depicted in (b). In the figures throughout this paper, regions of high energy are depicted by bright colors, whereas regions of lower energy are darker. In (c) and (d), classification results obtained during the segmentation procedure described in Sec. 3 are shown: MFCC-features are fed into a GMM to obtain a classification for each feature value and hence a detection curve (c). As may be observed, the classification curve is significantly fluctuating which is due to the high MFCC sampling rate in combination with the relatively high short-time variability in certain components of human

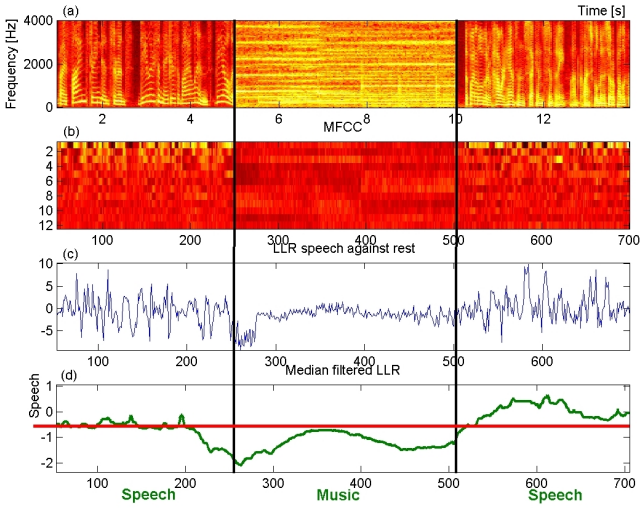

Figure 1: (a) Excerpt of a radio programme (14 seconds) consisting of music and speech segments. (b) Extracted MFCC-features. (c) The speech likelihood is detected using a MFCC-based GMM-classifier. (d) The results are subsequently smoothed by median filtering (green) and thresholded (red line) to obtain segments of the speech class *C2*.

speech. In order to obtain a more stable classification, a subsequent smoothing step is applied using a sliding median filter (green curve, (d)) which is followed by a threshold-based classification into speech segments (red line, (d)). In our example, segments exceeding the experimentally found threshold (i.e, values *above* the red line) are classified as speech. Although the smoothing has some of the desired effect of reducing fluctuations, it blurs the segment boundaries, resulting in an inexact segmentation. Furthermore, some of the fluctuations are still present resulting in an erroneous classification in the left speech segment.

A potential source of the classification errors illustrated before is that the smoothing operation is performed on the classification results and hence does not account for the properties of the actual signal features in the region of the smoothing window. From those considerations and inspired by a related approach using chroma features [4], this papers proposes to perform the smoothing at an earlier stage and incorporate this operation into the computation of the MFCC features. More precisely, we consider the spectral signal representation that is obtained by mel filtering the original signal

and compute certain short-time statistics of the mel spectrum coefficients followed by downsampling. Afterwards, the remaining part of the MFCC computation is performed, resulting in the so called *MFCC-ENS* (MFCC-Energy Normalized Statistics) features. In this, we are able to adjust the resulting feature resolution and sampling rate by suitably choosing the length of the statistics window and a downsampling factor.

Using the above segmentation scenario, we provide a comparison of the proposed MFCC-ENS features and the classical MFCC features. It turns out, that the MFCC-ENS are suitable to locally summarize the (MFCC-) audio properties. As a result, the MFCC-ENS-based classifiers yield less segmentation errors and more stable segmentation results than the standard MFCC do. We furthermore illustrate that the MFCC-ENS result from the MFCCs using a kind of seamless smoothing operation with the MFCCs at one end, which makes them rather promising for future applications.

The paper is organized as follows. In Sec. 2 we give the construction of the MFCC-ENS features and motivate it by the derivation of CENS-features (Chroma Energy Normalized Statistics) from chroma features as proposed in [4]. As an application, Sec. 3 details the segmentation scenario described above. Sec. 4 presents the evaluation results on both the segmentation performance and the comparison of MFCC-ENS and MFCCs. References to previous work will be given in the respective sections.

2. CONSTRUCTION OF MFCC-ENS FEATURES

To introduce the newly proposed features, we first summarize the standard process of computing MFCCs (2.1). To motivate the subsequently described approach to construct MFCC-ENS by using short-time MFCC statistics (2.3), we first briefly review the related approach of deriving CENS-from chroma features (2.2).

2.1 MFCCs

To compute MFCCs, successive blocks of an input signal are analyzed using a short time Fourier transform (STFT). For this, a typical block-length of 20 ms and a step size of 10 ms are used. For each of those temporal blocks, a feature vector is obtained as follows from its STFT-spectrum. First, the logarithmic amplitude spectrum is computed to account for the characteristics of human loudness sensation. To restrict the features to the human frequency range, only values $X(1),\ldots,X(N)$ corresponding to the region of $R = [133, 6855]$ Hz are used subsequently. In the next step, 40 frequency centers $f_1,\ldots,f_{40}$ are selected from R following a logarithmic scheme that approximates the Mel-scale of human frequency perception [9]. Using triangular windows Δ_i centered at the frequency centers f_i, a rough spectral smoothing is performed yielding 40 mel-scale components $M(i) = \sum_{j\in\Delta_i} \Delta_i(j)\cdot|X(j)|$, $1 \le i \le 40$. To approximately decorrelate the vector $(M(1),\ldots,M(40))$ a discrete cosine transform (DCT) is applied yielding $m = \mathrm{DCT}\cdot M$. As a last step, only the first 12 coefficients $m^{12} = (m(1),\ldots,m(12))$ remain, the other are rejected. We refer to [8] for more details on MFCCs.

As an Example, the top part of Fig. 2 shows MFCC features extracted from about 30 seconds of an audio signal containing three subsequent segments of orchestra music, male speech and a radio jingle comprising two speakers with background music.

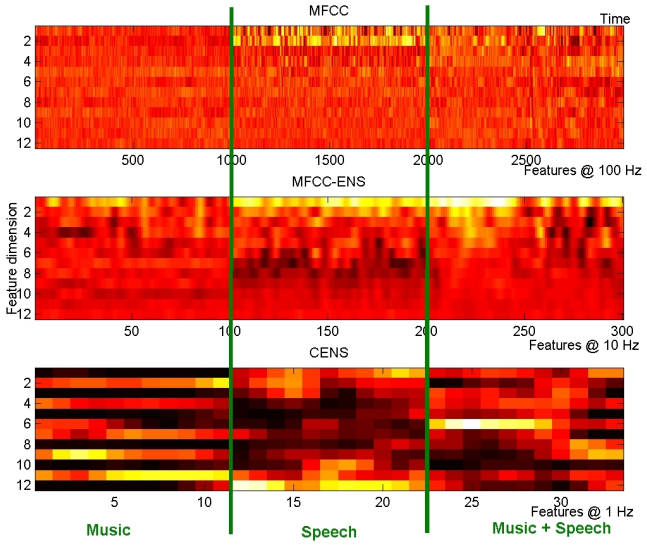

Figure 2: Three feature sets, MFCCs (top), MFCC-ENS$_{10}^{800}$ (center), CENS (bottom), extracted from an artificially concatenated audio fragment (33 seconds) consisting of orchestra music (left), male speech (center) and an radio jingle with two speakers and background music (right).

In speech processing applications one usually includes first and second order differences of m^{12} and the subsequent MFCC vectors to model temporal evolution. Those are also called *delta-* and *delta-delta-* coefficients. By furthermore including a single component to the initial 12 dimensions to represent the local signal energy, this results in a 39-component MFCC vector that is frequently used in speech recognition. Note that although we also considered delta- and delta-delta-coefficients for the applications discussed in the remainder of this paper, we will w.l.o.g. restrict our presentation to the basic 12-dimensional MFCC components in order to better illustrate the underlying principles.

2.2 Review of CENS features

Chroma-based audio features have turned out to be a powerful feature representation in the music retrieval context, where the chroma correspond to the twelve traditional pitch classes $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}$ of the equal-tempered scale, see [1]. To construct chroma features, the audio signal is converted into a sequence of twelve-dimensional chroma vectors. Let $v = (v(1), v(2)\ldots, v(12)) \in \mathbb{R}^{12}$ denote such a vector, then each entry expresses the short-time energy content of the signal in the respective chroma, where $v(1)$ corresponds to chroma C, $v(2)$ to chroma $\mathrm{C}^\sharp$, and so on. Such a chroma decomposition can be obtained for example by suitably pooling the spectral coefficients obtained from an STFT [1] as it is used for the MFCCs. Due to the octave equivalence, chroma features show a high degree of robustness to variations in timbre and instrumentation. A typical feature resolution is 10 Hz where each chroma vector corresponds to a temporal window of 200 ms.

To obtain features that robustly represent the harmonic progression of a piece of music, the computation of local statistics has been proposed in [4]. To absorb variations in

dynamics, in a preliminary step each chroma vector v is replaced by its relative energy distribution $v/\sum_{i=1}^{12} v(i)$. Vectors with insignificant energies are replaced by the uniform distribution. Afterwards, two types of short-time statistics are computed from these energy distributions. First, each chroma energy distribution vector $v = (v(1),\ldots,v(12)) \in [0,1]^{12}$ is quantized by applying a discrete 5-step quantizer Q yielding $Q(v) := (Q(v(1)),\ldots,Q(v(12))) \in \{0,\ldots,4\}^{12}$. The thresholds are chosen roughly logarithmic to account for the logarithmic sensation of sound intensity, see [9]. In second step, the sequence of quantized chroma distribution vectors is convolved component-wise with a Hann window of length $\mathbf{w} \in \mathbb{N}$ and then downsampled by a factor of $\mathbf{d} \in \mathbb{N}$. This results in a sequence of 12-dimensional vectors, which are finally normalized with respect to the Euclidean norm. The resulting features are referred to as CENS (chroma energy normalized statistics), which represent a kind of weighted statistics of the energy distribution over a window of $\mathbf{w}$ consecutive vectors. A configuration that has been successfully used for the audio matching taks, $\mathbf{w} = 44$ and $\mathbf{d} = 10$, results in a temporal resolution of 1 Hz [4]. The combination of *different* resolution levels has been successfully applied to obain multiresolution techniques for audio alignment [5].

In the bottom part of Fig. 2, the harmonic content of the orchestra music (first 10 seconds) is clearly visible in the CENS features which only contain significant energy in the chroma bands corresponding to the harmonics (comb-like structure). Also the harmonic content of the jingle (last seconds) is well-reflected by the characteristic comb structure. Due to the use of short-time statistics, the CENS reflect the coarse harmonic structure with smoothed-out local fluctuations.

2.3 MFCC-ENS-Construction

The basic approach to construct smoothed MFCCs consists of applying the short-time statistics operations from the CENS construction at a suitable instant during the MFCC computation. To include all aspects of the MFCCs which are related to human perception into the short-time statistics, the MFCC-ENS computation starts using the mel-scale coefficients $M = (M(1),\ldots,M(40))$. Subsequently, the following steps are performed:

- M is replaced by a normalized version $M/\sum_{i=1}^{40} |M(i)|$ in order to achieve invariance w.r.t dynamics. If $\sum_{i=1}^{40} |M(i)|$ is below a threshold, M is replaced by the uniform distibution.
- Each component of the resulting vector is quantized using the above discrete quantizer $Q : [0,1] \to \{0,1,2,3,4\}$ which is more precisely defined by $Q(a) := 0$ for $a \in [0,0.05)$, $Q(a) := 1$ for $a \in [0.05,0.1)$, $Q(a) := 2$ for $a \in [0.1,0.2)$, $Q(a) := 3$ for $a \in [0.2,0.4)$, and $Q(a) := 4$ for $a \in [0.4,1]$. As a result, besides the rough log-characteristics, only the more significant components are preserved and reduced into four classes. This step performs a kind of *frequency statistics*.
- To furthermore introduce *time*-based statistics, the resulting sequence of quantized 40-dimensional vectors is smoothed by filtering each of the 40 components using a Hann-window of length ℓ ms.
- As a last step, the vector sequence is downsampled by an integer factor resulting in a vector sequence of sampling rate f Hz.
- Each vector is then decorrelated using a DCT operation as performed at the end of the MFCC computation. By restriction to the lowest 12 coefficients of each DCT-vector, we obtain a vector sequence MFCC-ENS_f^ℓ of smoothed MFCCs with a smoothing range of ℓ ms and sampling rate of f Hz.

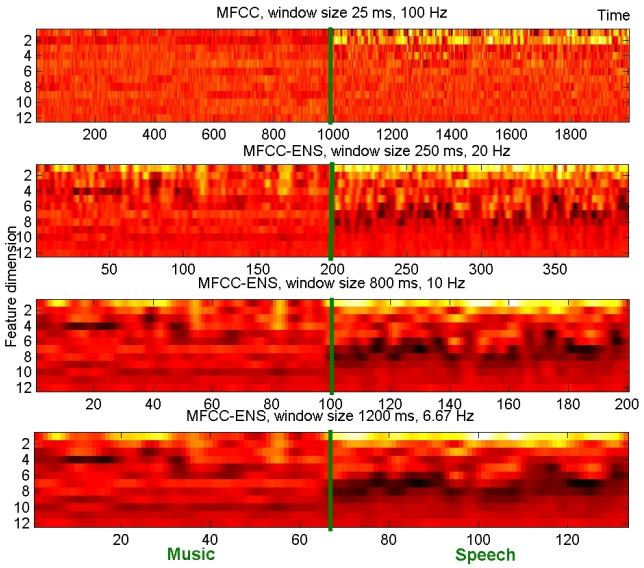

Figure 3: Evolution of MFCC-ENS for different parameters. From top to bottom: MFCCs and feature sets $\text{MFCC-ENS}_{20}^{250}$, $\text{MFCC-ENS}_{10}^{800}$, $\text{MFCC-ENS}_{6.67}^{1200}$ for the frist 22 seconds (music and speech) of the audio example shown in Fig. 2.

By construction, the MFCC-CENS's time resolution may be flexibly chosen by adjusting the window sizes and downsampling factors which are directly related to the quantities ℓ and f. As an example, the center part of Fig. 2, shows $\text{MFCC-ENS}_{10}^{800}$ (a window length equivalent to 800 ms at a feature sampling rate of 10 Hz) for the given audio example.

As the DCT is a linear mapping, the smoothing operation that is performed during MFCC-ENS computation in the mel-spectral domain also takes effect after applying DCT. As an illustration, Fig. 3 compares the classic MFCC features (top) to the features obtained by the gradual transition from $\text{MFCC-ENS}_{20}^{250}$ to $\text{MFCC-ENS}_{6.67}^{1200}$.

We note that one particular parameter in the MFCC-ENS computation that may be adjusted in the future is the quantizer Q that, to this point, has been copied from the CENS computation. Because MFCCs are already based on a logarithmic amplitude spectrum, a different choice of Q might be more appropriate. As, however, replacing Q by a linear quantizer did not result in a better performace during our segmentation tests, a more detailed investigation was postponed.

Transform-domain filtering has long been used to obtain robust feature representations for speech processing. An important step was the introduction of the RASTA processing concept [3] that was used to suppress log-spectrum components by applying recursive bandpass filterbanks to the

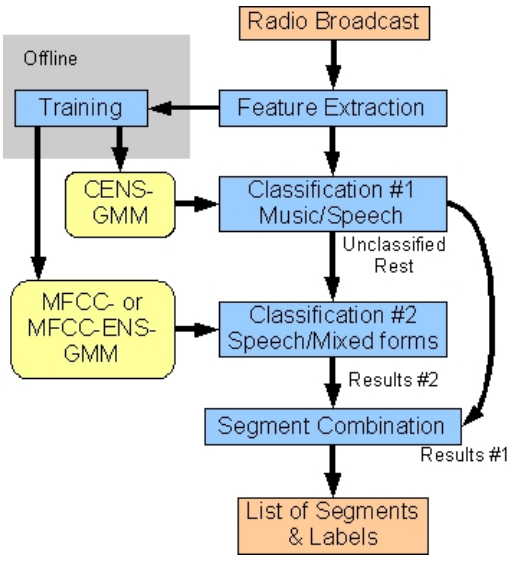

Figure 4: Overview on the two-stage segmentation procedure.

spectral trajectories, thereby averaging out components that change at higher or lower rates than perceivable by humans. While RASTA processing and related techniques have been successfully applied to noise suppression and speech enhancement, our approach puts an additional focus on an adjustable feature resolution and resulting data rate, which is of importance for the targeted speech retrieval tasks.

3. APPLICATION TO SPEECH SEGMENTATION

As an application, we consider the segmentation scenario described in the introduction. In particular, we consider the task of segmenting broadcast radio programmes where the possible classes are *Music (C1)*, *Speech (C2)* and *Speech+Music (C3)*. Fig. 4 shows an overview of our two-stage segmentation procedure consisting of an offline training phase and an online segmentation phase.

In the *training phase*, a suitable amount of audio material is recorded, manually segmented and labeled using the classes *(C1)-(C3)*. Note that for practical purposes, class *(C3)* was choosen to also subsume audio effects and other types of noise that could not always be properly separated from the other classes. Hence, a more proper label for class *(C3)* will be *Mixed forms*. For each class, an equal amount of audio data is gathered and both MFCC- and CENS-features are extracted at specific sampling rates (that generally differ from MFCC to CENS), resulting in six feature sets $\mathscr{F}_1^{\mathrm{MFCC}}$ - $\mathscr{F}_3^{\mathrm{MFCC}}$ and $\mathscr{F}_1^{\mathrm{CENS}}$ - $\mathscr{F}_3^{\mathrm{CENS}}$. For each of those feature sets, a Gaussian mixture model (GMM) is trained which is used in the subsequent segmentation phase.

During the (online) *segmentation phase*, sequences of both MFCC- and CENS-features are extracted from a recorded audio signal at the same sampling rates as used during training. Subsequently, two GMM-based classifiers are used for classification. The first classifier works on the extracted CENS-features and uses the CENS-based GMMs to perform a binary classification into the two classes *Music* and *Non-Music*. In our settings it turns out that a log-likelihood ratio test based on the GMMs for speech and music is a good approximation for this task. The segments classified as *Music* are labeled by *(C1)* and are used for the later on segment generation. The remaining segments are handed over to the second classifier. This classifier uses the MFCC-trained GMMs to perform a binary classification into the classes *Speech* and *Mixed forms*. For this, a log-likelihood ratio test using the GMMs for the classes music and mixed forms is used. Segments classified as speech are labeled as *(C2)* while the mixed forms results are labeled *(C3)*. The subsequent step of segment combination assembles the outputs of both classifiers and outputs a properly formated list of labeled segments. The overall system will be called *MFCC-based* segmenter.

For use with the MFCC-ENS features, the MFCCs in the above procedure are replaced by the MFCC-ENSs. For example, the MFCC-training sets are replaced by $\mathscr{F}_1^{\mathrm{MFCC-ENS}}$ - $\mathscr{F}_3^{\mathrm{MFCC-ENS}}$ for a suitably chosen MFCC-ENS-resolution. While the other components of the segmenter stay the same, the resulting system will be called *MFCC-ENS-based* segmenter.

We note that the above GMM-based classifiers output classification likelihoods at a sampling rate induced by the feature sequence. To obtain a stable classification output, a subsequent smoothing operation based on median filtering followed by a threshold-based decision as illustrated in the introduction is performed which depends on the actual feature resolution and feature type. Note that the thresholds used in our evaluations have been determined experimentally based on our training corpus.

The basic strategies used in the latter approach to audio segmentation have been proposed and investigated in several previous studies. A combined use of MFCC- and chroma-based features to account for the particularities of both speech and music was recently described in an approach to speech/music discrimination [7]. Among various other classification strategies, GMMs have been widely used in the audio domain. An application to discriminating speech and music is for example described in [2].

4. EVALUATION

To illustrate the effect of MFCC-ENS-based smoothing, Fig. 5 revisits the audio fragment shown in Fig. 1. Parts (b)-(d) of the figure show the corresponding results for speech detection obtained using the MFCC-ENS$_{10}^{800}$ features instead of MFCCs. For the subsequent median filtering, the window size was adapted in order to obtain equivalent temporal smoothing regions with both approaches. It can be observed that the MFCC-ENS-based detection is more stable und short-term fluctuations are clearly reduced. As a result, the left hand speech segment, which was wrongly classified using MFCCs is now classified correctly.

For a larger-scale comparison of the segmentation performance, we prepared an audio collection consisting of the following material taken from a classic radio station. For training the MFCC- and CENS-based GMMs, we used 20 minutes of audio for each of the three classes *(C1)*, *(C2)* and *(C3)*. For training, we used the Expectation Maximization algorithm which was run until convergence. The GMMs consisted of 16 mixtures each with dimensions of 12 (CENS) and 39 (MFCCs). For the MFCC-ENS-based segmenter we used MFCC-ENS$_{10}^{800}$ features. The training set was increased

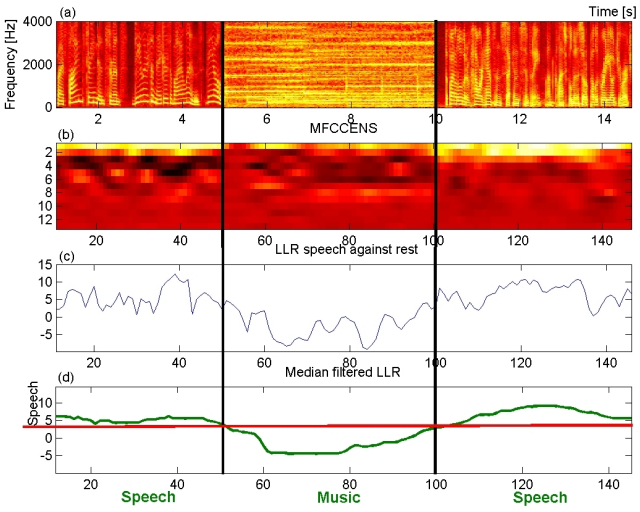

Figure 5: (a) Audio example revisited from Fig. 1. (b) Extracted MFCC-ENS$_{10}^{800}$ features. (c) Log-likelihood ratio of speech against mixed forms class. (d) Log-likelihood (green) smoothed by median-filtering (length 20 samples) with speech detection threshold (red).

Table 1: Confusion matrix for results of MFCC-based segmenter (left) and MFCC-ENS-based segmenter (right). Used classes: *Music (C1)*, *Speech (C2)* and *Mixed forms (C3)*.

Seg. result [%] ↓	**MFCC** True class			**MFCC-ENS** True class		
	C1	*C2*	*C3*	*C1*	*C2*	*C3*
C1	98.3	0	0	98.3	0	0
C2	1.7	84.68	33.89	1.7	97.55	6.19
C3	0	15.32	66.11	0	2.45	93.81

to 40 minutes (speech) and 100 minutes (mixed foms) in order to account for the lower feature resolution.

The segmentation was performed using the procedure described in Sect. 3. Our test data consisted of 4:09 hours of a contiguous audio programme recorded from the radio station and labeled manually. The material comprises 206.42 minutes of music *(C1)*, 13.5 minutes of speech *(C2)* and 30.15 minutes of *(C3)*-segments (mainly jingles and commercials consisting of mixed speech and music). For this data, the overall rate of correct classifications using the MFCC-based segmenter was 93.68%, where we evaluated one classification result per second. The left part of Table 1 shows the confusion matrix for the three involved classes. As might be expected, the class *C3* containing superpositions of music and spoken language causes the largest classification errors.

The right part of Table 1 shows the corresponding confusion matrix for the MFCC-ENS-based segmenter. As may be observed, confusion of classes *C2* and *C3* is significantly reduced due to the improved MFCC-ENS-based classifier. The overall rate of correct classifications is 97.72%. A manual inspection of the log-likelihood curves used for segmentation confirms the observation that speech segments are now much more clearly separated from the other classes as was already illustrated in Fig. 5.

We conclude this section by remarking that although the size of the training set in minutes was larger when using MFCC-ENS our tests indicate that a further increase may be beneficial. This will we subject of future investigations.

5. CONCLUSIONS

In this paper, we introduced a class of audio feature, MFCC-ENS, which is constructed by computing suitable short-time statistics of the well-known CENS-feature. More precisely, quantization and smoothing operations are performed on the mel-spectrum representation to generate compact summaries of a signal's short-time acoustic contents. By introducing parameters controlling the new features' time resolution, the feature granularity may be flexibly adjusted with the standard MFCCs resolution appearing as a special case. The features were evaluated for the application of segmenting broadcast radio. It was shown that due to the smoothing properties, MFCC-ENS can aid in overcoming unstable segmentation as may result when using MFCCs.

Future work will deal with further investigating MFCC-ENS and their properties. Innovative applications using MFCCs such as unsupervised discovery of speech patterns [6] that right now rely on performing temporal smoothing in a higher level step may also benefit from the proposed MFCC-ENS features.

REFERENCES

[1] M. A. Bartsch and G. H. Wakefield. Audio Thumbnailing of Popular Music Using Chroma-based Representations. *IEEE Trans. on Multimedia*, 7(1):96–104, Feb. 2005.

[2] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proc. ICASSP 1999, Phoenix, USA*, pages 1432–1435, 1999.

[3] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.

[4] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, February 2008.

[5] M. Müller, H. Mattes, and F. Kurth. An Efficient Multiscale Approach to Audio Synchronization. In *ISMIR, Victoria, CND*, 2006.

[6] A. S. Park and J. R. Glass. Unsupervised Pattern Discovery in Speech. *IEEE Trans. on Audio, Speech, and Language Processing*, 16(1):186–197, Jan. 2008.

[7] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis. A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks. *IEEE Trans. on Multimedia*, 10(5):846–857, Aug. 2008.

[8] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer Verlag, 1990.