Histogram Equalization-Based Features for Speech, Music, and Song Discrimination

Ascensión Gallardo-Antolín and Juan M. Montero, *Member, IEEE*

***Abstract*—In this letter, we present a new class of segment-based features for speech, music and song discrimination. These features, called PHEQ (Polynomial-Fit Histogram Equalization), are derived from the nonlinear relationship between the short-term feature distributions computed at segment level and a reference distribution. Results show that PHEQ characteristics outperform short-term features such as Mel Frequency Cepstrum Coefficients (MFCC) and conventional segment-based ones such as MFCC mean and variance. Furthermore, the combination of short-term and PHEQ features significantly improves the performance of the whole system.**

***Index Terms*—Acoustic features, audio classification, HEQ-based features, parameterization, speech/music/song discrimination.**

I. Introduction

THE outstanding success of Internet and the rapid growth of multimedia contents on it has brought up interest for developing techniques to automatically classify these contents. In particular, audio files and audio tracks in videos contain relevant information about the nature and content of the multimedia file. In this context, audio classification plays an important role as preprocessing step in a variety of more complex systems related to information retrieval in music or multimedia content extraction, annotation and indexing.

Previous work in the area of audio classification has focused mostly in audio event classification [1] and speech/music discrimination [2]–[4]. In this letter we investigate the problem of automatically classifying collections of audio files in three acoustic classes: speech, instrumental music and song (music with singing voice). These categories can be seen as a broad classification of data previous to refined classification or further processing.

The problem proposed in this letter is very closely related to speech/music classification tasks. However, speech, music and song discrimination is more complex and the performance of such kind of systems is expected to be lower especially due to the strong correlation between the singing voice and the corresponding background music [5] and the similarities between nonsinging and singing voices. In fact, it has been observed that audio files containing music with a large speech content (opera, rap, ...) tend to be classified as speech [2]. In addition, the large amount of different styles of music and song deteriorates the performance of these systems. In spite of these differences, similar approaches for feature extraction and classification have been utilized for both tasks.

One of the main issues addressed in the literature of this field is the study of audio descriptors more suitable for classification. Among the different kind of features proposed for speech/music discrimination, it is worth mentioning the well-known Mel-Frequency Cepstral Coefficients (MFCC) [4], [6], Line Spectral Frequencies (LSF) [2], Zero-Crossing Rate (ZCR) and Frame Energy [4], [6] and more specific parameters such as Spectral Centroid, Spectral Flux, Spectral Rolloff [6] or Chroma-Vector based features [4].

Most of the previously mentioned features are *short-term* characteristics in the sense that they are extracted on a frame-by-frame basis (typically, the frame period used for speech/audio analysis is about 10–20 ms). In some works, they are directly fed to the classifier [2], [5]; however, in other cases, it is preferred to consider the descriptors information over several consecutive frames. For doing that, short-term features are mapped to their statistics computed over a certain window. This procedure makes sense because in contrast to other tasks (like speech recognition), it is reasonable to assume that music, speech or song lasts at least 1–2 s. We refer to these characteristics as *segment-based* features.

The most common segment-based features are the mean and variance of a full segment [7], although higher order statistics such as kurtosis or skewness have been reported [6] or complex combinations of all these features (as, for example, the quotient between the maximum value of the feature in the segment and the mean [4]). The problem is that is not easy to determine *a priori* which kind of statistics of a certain parameter are more suitable for the classification task. In practice, a certain statistic is chosen empirically after an exhaustive experimentation [4], [6] or by using feature selection techniques [7]. However, both methods are computationally expensive.

In this letter, we have focused in the development of new segment-based features for speech, music and song discrimination. They are derived from the nonlinear relationship between a reference distribution and the short-term feature distributions computed at segment level. Also, we show that the combination of

A. Gallardo-Antolín is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain (e-mail: gallardo@tsc.uc3m.es).

J. M. Montero is with the Speech Technology Group, Department of Electronic Engineering, Universidad Politécnica de Madrid, Madrid, Spain (e-mail: juancho@die.upm.es).

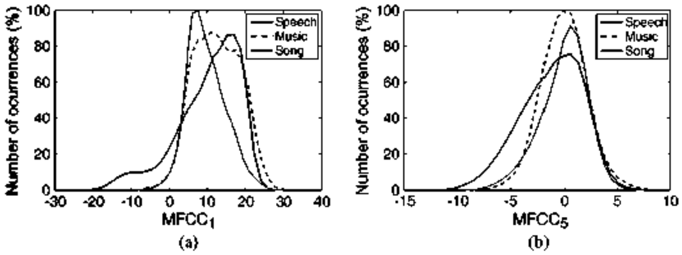

Fig. 1. Histograms of the (a) first ($MFCC_1$) and (b) fifth ($MFCC_5$) MFC coefficient for speech, music, and song.

short-term features and the proposed ones can improve the performance of the system.

This letter is organized as follows. Section II motivates the use of the proposed segment-based features for audio classification. Section III describes the extraction procedure. Section IV presents the experiments and discusses the results. Finally, some conclusions are drawn in Section V.

II. Segment-Based MFCC Features of Audio Signals

In this work, we consider the well-known MFCC parameters as short-term features, although the procedure described in this letter could be applied to other types of audio descriptors.

Once the MFCC are extracted each 10–20 ms, a common choice for the segment-based features is the MFCC mean and/or variance computed over windows of 1–2 s length. However, in general, the mean and variance are not discriminative enough for distinguishing between the different audio classes. This observation is illustrated in the following example. Fig. 1 represents the probability distributions of the first and fifth MFCC for the three acoustic classes considered. They have been estimated by smoothing the histograms obtained from a subset of the database described in Section IV. From Fig. 1, it is clear that the distributions of music, song and speech not only differ on the mean and the variance, but also in the shape. Therefore, it seems necessary to include the information regarding higher order moments into the parametric representation of each audio segment. This fact has already been highlighted by other authors in which the kurtosis and skewness are also considered [6].

In these conventional approaches, each one of the moments considered must be computed explicitly and an extensive experimentation must be performed in order to determine which of them are more relevant for the audio classification task. For circumventing this problem, we propose a new parameterization of audio signals which implicitly takes into account the mean, standard deviation and shape of the acoustic features belonging to each audio segment, as described in the next section.

III. Histogram Equalization-Based Segmental Features for Audio Classification

The proposed segment-based features, called PHEQ (Polynomial-Fit Histogram Equalization), consist of a set of parameters derived from the nonlinear transformation function which performs a mapping between the feature probability distribution of each audio segment and a reference one. Specifically, the PHEQ features are the coefficients of the polynomial approximation of these transformation functions which are determined on a segment-by-segment basis by using Histogram Equalization (HEQ)-based techniques.

A. *Histogram Equalization (HEQ)-Based Transformations*

Recently, HEQ have been widely applied for compensating the nonlinear distortions due to the presence of noise in automatic speech recognition systems [8], [9]. For our purposes, HEQ will be used for determining the transformation which converts the distribution of the acoustic vectors contained in an audio segment into a predefined one. As target distribution we have chosen the simplest one: a single-mode Gaussian with zero mean and unit variance. The HEQ-based transformation provides the classification algorithm with information regarding the mean, the variance and other higher order moments of the feature distributions [8].

For finding the transformation we have followed the procedure described in [8]. The theory behind HEQ establishes that given two random variables x and z, whose probability density functions are respectively, $p_x(x)$ and $p_z(z)$, the transformation $z = T\{x\}$ which converts $p_x(x)$ into $p_z(z)$ can be obtained as

$$z = T\{x\} = C_z^{-1}\left[C_x(x)\right] \tag{1}$$

in which $C_x(.)$ and $C_z^{-1}(.)$ represents the cumulative density function (CDF) of x and the inverse function of the cumulative probability of z, respectively.

In practice, z is the reference distribution, so C_z^{-1} is known because it corresponds to the inverse function of the CDF of a single-mode Gaussian with zero mean and unit variance. In addition, $x = \{x_1, \cdots, x_N\}$ corresponds to the feature vectors in an audio segment of length N, so its CDF must be estimated for each segment. For that, we have followed the approach named OSEQ (order statistics-based equalization) described in [8] in which the CDF of x is estimated using the following formulation,

$$C_x(x_n) = \frac{r(x_n) - 0.5}{N}, \qquad n = 1, \cdots, N \tag{2}$$

in which $r(x_n)$ denotes the rank of x_n when the elements of x are sorted in ascending order. Using (1) and (2), the value of the reference distribution z_n corresponding to x_n is computed as follows

$$z_n = T\{x_n\} = C_z^{-1}\left[C_x(x_n)\right] = C_z^{-1}\left[\frac{r(x_n) - 0.5}{N}\right]. \tag{3}$$

By applying this formula to each of the N elements of the audio segment x, N pair of points $\{x_n, z_n\}$, $n = 1, \cdots, N$ of the nonlinear transformation $T(x)$ are obtained. A parametric version of $T(x)$ is computed by using the procedure described in Section III-B.

B. *PHEQ Segment-Based Features*

The parametric version of HEQ we have implemented is similar to the so-called PHEQ ("*Polynomial-Fit Histogram Equalization*") [9] approach, in the sense that the nonlinear transformation is approximated by a Kth order polynomial. One of the main differences is that in [9] the estimation of the CDF is performed through a quantile-based method and in our case, a direct estimation of the CDF is performed as described in

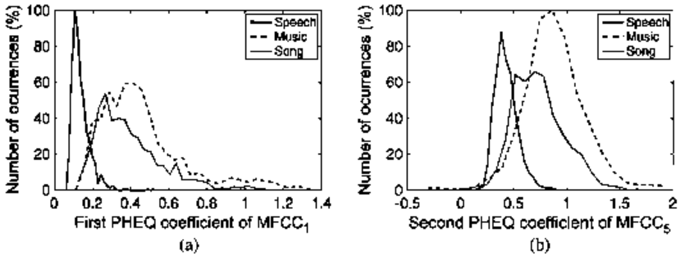

Fig. 2. Histograms of the (a) first PHEQ coefficient of $MFCC_1$ and (b) second PHEQ coefficient of $MFCC_5$ for speech, music, and song.

Section III-A. The coefficients of the polynomial approximation are PHEQ features which are the input of the classifier.

The extraction of the PHEQ features is performed in the following steps.

1) Extract the frame-by-frame acoustic features (in this case, MFCC).
2) Define a window length for segment processing (in our case, 1.5 s). For each component of the feature vectors contained in the current audio segment:
   - Sort the data of the segment in ascending order.
   - Compute the corresponding transformed features using the OSEQ approach as expressed in (3).
     At the end of this step, a set of pair of points (x_n, z_n) are obtained in which each pair represents a feature component and its transformed counterpart.
   - Approximate the nonlinear transformation defined by the sets of points computed in the previous step by a Kth order polynomial $p(x)$ defined by

     $$z = T(x) \approx p(x) = \sum_{k=0}^{K} a_k x^k. \tag{4}$$

     The polynomial coefficients $\{a_k\}$ are computed by using a classical least-square approach from the set of points [9]. These are the PHEQ features for the current audio segment.
3) Repeat the process in step 2 for all the segments in the audio sequence.

As the PHEQ transformation indicates somehow the degree of similarity between the feature distribution and a Gaussian distribution with zero mean and unit variance, the coefficients of its approximation will contain the information regarding the mean, variance, and, in general, shape of the distribution. In fact, if the transformation is approximated by a zero order polynomial, the zero order coefficient is directly the mean of the distribution. If a first order polynomial is considered, the zero order coefficient is the quotient between the mean and the standard deviation and the first order one is the inverse of the standard deviation. Higher order moments are related to the polynomial coefficients if higher order polynomial approximations are used.

An example of the discrimination capability of the PHEQ features is shown in Fig. 2(a) and (b), which represent, respectively, the histograms of the first PHEQ coefficient of the first MFCC and the second PHEQ coefficient of the fifth MFCC for speech, music and song. It can be observed that PHEQ features can be used for discriminating between the three acoustic classes considered.

TABLE I
CLASSIFICATION ACCURACY FOR DIFFERENT ACOUSTIC FEATURES

Feature	Classification Accuracy (%)
MFCC	72.34%
Mean	62.46%
Variance	73.24%
Skewness	56.43%
Kurtosis	47.17%
PHEQ (order 1)	74.71%
PHEQ (order 2)	75.29%
PHEQ (order 3)	71.80%
PHEQ (order 4)	68.77%

IV. RESULTS

A. Experimental Protocol and Database

According to our knowledge there is no a common database for this task, we have created a specific one. This database comprises 800 excerpts of speech, 901 of instrumental music, and 739 of songs (singing voice) yielding to a total of 2440 audio files. They were recorded from mp3 and cds with a 22.05 KHz sampling rate, covering a wide variety of speakers and musical genres (classical, rock, pop, etc.). As special care was taken for not mixing instrumental music and song, when needed, audio files were manually segmented and relabelled.

Since this database is too small to achieve reliable classification results, we have used a 6-fold cross validation to artificially extend it, averaging the results afterwards. Specifically, we have split the database into six disjoint balanced groups. One different group is kept for testing in each fold, while the remainder are used for training.

In the experiments, we have used a GMM-based classifier developed using the HTK package [10]. For modelling each one of the acoustic classes, after a preliminary experimentation, we chose GMMs with 256 and eight gaussians when using, respectively, the short-term and the segment-based features.

B. Results With Short-Term and Segment-Based Features

For the computation of the short-term features, the audio signal was analyzed with a Hamming window of 25 ms length and 12 MFCC were extracted at a frame period of 10 ms. The segment-based characteristics (mean, variance, skewness, kurtosis and PHEQ) were calculated from the MFCC over segments of 1.5 s length with overlap of 0.5 s.

Table I shows the results achieved in terms of classification accuracy (percentage of files correctly classified). For the case of PHEQ, we have tried polynomial approximations from order 1 to 4. Note that with a polynomial approximation of order Kth the number of PHEQ coefficients is $K+1$. Results with a zero order approximation are not explicitly included because this case is equivalent to the MFCC mean (third row).

As can be observed, using only the mean, skewness or kurtosis does not outperform the short-term features. However, the variance produces similar classification accuracies with MFCC. With respect to PHEQ features, the results obtained with first and second order approximations are better than those obtained with short-term features, mean and variance. In particular, the improvement achieved by PHEQ (order 2) with respect to MFCC is statistically significant according to the

TABLE II
CONFUSION MATRICES FOR THE MFCC-BASED AND PHEQ-BASED SYSTEMS

	MFCC			PHEQ (order 2)		
Class	Speech	Music	Song	Speech	Music	Song
Speech	96.00%	0.38%	3.62%	93.62%	0.38%	6.00%
Music	8.00%	52.83%	39.17%	2.22%	62.82%	34.96%
Song	15.16%	14.34%	70.50%	3.65%	25.71%	70.64%

confidence intervals calculated for a confidence of 95% (see [11], for details). On the contrary, higher-order approximations do not outperform MFCC and variance. This can be explained because of overfitting problems when increasing the order of the polynomials.

C. Results With the Combination of Short-Term and PHEQ Features

As short-term features implicitly carry local information about the sequential nature of audio signals, they may be complimentary to the information provided by the segment-based ones. To gain insight into this possibility, we have analyzed the confusion matrices produced by the MFCC-based and PHEQ-based (order 2) systems. Table II shows the corresponding confusion matrices. In this table, the rows correspond to the correct class, the columns to the hypothesized one and the values in it are computed as the average for the six folds. As can be observed, in both systems speech is the less confusable class whereas music is the highest confusable one. In the MFCC-based system, 8% and 39.17% of music files are classified as speech and as song, respectively. However, in the PHEQ-based system, music is better recognized and the errors are mainly due to confusions with song (34.96%). The song class presents a different pattern of behaviour in each system: in the MFCC-based one it is confused with almost the same percentage with speech (15.16%) and music (14.34%); however in the PHEQ-based one, it is mainly confused with music (25.71%).

The analysis of the confusion matrices allows us to conclude that both systems produce different kind of errors, so they are suitable candidates for combination. To corroborate this observation, we have carried out several experiments in which the combination of short-term and PHEQ features is performed at segment level using the well-known product rule [12].

Table III shows a summary of the main results obtained. "Normalized MFCC" and "Normalized MFCC-D" refer, respectively, to MFCC normalized with respect to their mean (CMN) and normalized MFCC augmented with their corresponding first derivatives. As can be observed, the combination of short-term and PHEQ features always outperform the performance of the individual systems. In addition, the differences are statistically significant. It is also worth mentioning that normalized MFCC produces slightly better results than MFCC whereas the combination of PHEQ with normalized MFCC significantly improves the results of just normalized MFCC. Finally, although the use of the first derivatives in other works in the literature did not provide a consistent improvement for music/speech discrimination tasks, in our study, the best results were obtained when also including these additional features.

TABLE III
RESULTS WITH THE COMBINATION OF SHORT-TERM AND PHEQ FEATURES

Feature	Classification Accuracy (%)
MFCC	72.34%
Normalized MFCC	73.36%
PHEQ (order 2)	75.29%
MFCC + PHEQ (order 2)	79.14%
Normalized MFCC + PHEQ (order 2)	81.52%
Normalized MFCC-D + PHEQ (order 2)	83.81%

V. CONCLUSIONS

In this letter, we have developed new segment-based features more appropriate to capture the differences between instrumental music, singing voice and speech. These PHEQ features are the coefficients of the polynomial approximations of the transformation functions between the short-term feature distributions computed at segment level and a zero-mean unity-variance Gaussian reference distribution. Results show that PHEQ characteristics obtained from second-order approximations are able to discriminate between the three acoustic classes considered and they outperform conventional features such as short-term MFCC or MFCC mean and variance. Furthermore, a combination of short-term MFCC and PHEQ features achieves better results than the individual systems.

REFERENCES

[1] A. Temko and C. Nadeu, "Classification of acoustic events using svm-based clustering schemes," *Pattern Recognit.*, vol. 39, no. 4, pp. 682–694, Apr. 2006.

[2] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proc. IEEE ICASSP*, 2000, pp. 2445–2448.

[3] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech Commun.*, vol. 40, no. 3, pp. 351–363, May 2003.

[4] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 846–857, Aug. 2008.

[5] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proc. IEEE ICASSP*, 2008, pp. 1885–1888.

[6] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A speech/music discriminator for radio recordings using bayesian networks," in *Proc. IEEE ICASSP*, 2006, pp. 809–812.

[7] B. Schuller, B. J. B. Schmitt, D. Arsic, S. Reiter, M. Lang, and G. Rigoll, "Feature selection and stacking for robust discrimination of speech, monophonic singing and polyphonic music," in *Proc. IEEE ICME*, 2005, pp. 840–843.

[8] J. C. Segura, C. Benítez, A. de la Torre, A. J. Rubio, and J. Ramírez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 517–520, May 2004.

[9] S.-H. Lin, Y.-M. Yeh, and B. Chen, "Exploiting polynomial-fit histogram equalization and temporal average for robust speech recognition," in *Proc. ICSLP*, 2006, pp. 2522–2525.

[10] S. Young *et al.*, *HTK-Hidden Markov Model Toolkit (Ver 3.2)*. Cambridge, MA: Cambridge Univ., 2002.

[11] N. A. Weiss and M. J. Hasset, *Introductory Statistics*. Reading, MA: Addison-Wesley, 1993, pp. 407–408.

[12] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Commun.*, vol. 37, pp. 303–319, 2002.