Proc. of the National Conference on Communications (NCC) 2008

Singing Voice Detection in North Indian Classical Music

Vishweshwara Rao, S. Ramakrishnan and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai 400076, India
Email: {vishu, ramakrishnan, prao}@ee.iitb.ac.in

***Abstract*— Singing voice detection is essential for content-based applications such as those involving melody extraction and singer identification. This article is concerned with the accurate detection of singing voice phrases in north Indian classical vocal music. The component sound sources in such music fit into a typical framework (voice, rhythm and drone). We have used this a-priori knowledge to enhance the voice in the presence of accompaniment. A Gaussian Mixture Model (GMM) classifier is evaluated using frame-level feature vectors extracted from a representative data set. A threshold based method, applied to suitable audio features, is then used to automatically divide the audio signals into variable-length homogenous segments i.e. vocal or non-vocal. Segment-level classification decisions are made by grouping frame-level decisions over individual segments. The performance of the classifier is evaluated in terms of the classification accuracies of the vocal and non-vocal frames.**

I. Introduction

The melody of a musical piece is a very important attribute from a Music Information Retrieval (MIR) perspective. The melody may be defined as the temporal trajectory of the pitch of the predominant melodic instrument which, in note-based music, corresponds to the sequence of note pitches and durations. Melody extraction therefore involves detecting the presence of the melodic instrument in the audio record, and tracking its pitch. In most Indian music, the predominant melodic instrument is the human singing voice. In classical vocal music, the voice is supported by instruments providing the drone (*tanpura*) and rhythmic accompaniment (*tabla*). Pitch detection in the presence of simultaneously playing instruments is a difficult problem. Methods that track the dominant pitch by applying smoothness constraints over continuously sung phrases achieve reasonable accuracies [1]. The reliable detection of singing voice segments is necessary for the effectiveness of such methods. It is also essential in systems that attempt to identify the singer/artist [2] [3].

Singing voice detection can be viewed as an audio classification problem where features that distinguish the vocal regions from pure instrumental regions are fed to a classifier previously trained on labeled data. In [4] a Multi-Layer Perceptron (MLP) is used to segment songs into vocal and non-vocal regions based on Perceptual Linear Prediction (PLP) features. In [2] features such as energy, average Zero Crossing Rate (ZCR), harmonic coefficients and spectral flux are computed at regular intervals and a threshold-based measure is used to detect the onset of singing. In [3] a combination of band-pass filtering and harmonicity detection are used to extract vocal segments with limited success. [5] uses a Support Vector Machine (SVM) classifier to detect the presence of the singing voice using features such as Linear Prediction Coefficients (LPC), LPC-derived Cepstrum (LPCC), MFCC spectral power, short-term energy and ZCR. [6] uses MFCCs along with a GMM classifier. Additionally, frame-level likelihoods for individual classes are pooled over automatically detected segments to arrive at a segment label. In [7] trained multi-model HMMs and a-priori musical knowledge are used to separate vocal and non-vocal regions.

In the present work we investigate audio features suitable for use in a simple rule-based classifier to detect singing segments in Indian classical vocal music where the singer is accompanied by the *tanpura* and *tabla*. The accompanying instruments exhibit certain characteristics similar to the singing voice necessitating a reconsideration of audio features for vocal/non-vocal classification for Indian classical music. In the interest of obtaining robust classification across a variety of audio signals under this category, we attempt to find features based on the study of known and observed acoustic signal characteristics.

In the next section, we discuss the signal characteristics of the singing voice and the accompaniment. A method to suppress the accompanying instrumental background relative to the singing voice is then proposed. For classification we follow the framework developed in [6]. We describe the features we have used as input into the classification module and provide frame-level classification results. Further, audio features that exhibit variation in value across vocal-instrumental boundaries or vice versa are described and evaluated for segment boundaries detection. The segments between the boundaries so obtained are assumed to be homogenous in terms of vocal/non-vocal categorization. Two methods of deciding the final segment labels are then described.

Table 1. Statistics of the vocal and non-vocal segments for the audio data set

Segments	Number	Longest	Shortest	Avg. duration	Total duration
Vocal	187	15.13 s	0.53 s	5.44 s	1018.01 s
Non-vocal	208	13.20 s	0.11 s	1.73 s	359.62 s

II. Signal Characteristics

A. Singing Voice

That the range of fundamental frequency (F0) for singing is greater than that for speech is well known. Additionally the singing voice, in north Indian classical singing, is replete with large pitch modulations, which serve an important aesthetic function within the melodic contour. These modulations are further magnified in the harmonic content of the voice located in the higher regions of the spectrum. However, the locations of significant voice harmonics in the spectrum are usually restricted to below 5 kHz.

With respect to articulation in singing, the vowel durations are much longer than in speech. In Indian classical singing especially, the ratio of vowel to consonant duration within a single sung phrase is very high, with the dominant vowel being /a/. The duration of sung phrases also shows large variation, from as short as 1 second to as long as 15 seconds.

B. Drone (Tanpura)

The *tanpura* is an overtone-rich stringed instrument, usually having four strings that are repeatedly plucked in succession throughout the music performance. Two of the strings are tuned to the singer's tonic, one to an octave above this tonic and the last one is usually tuned to either the fourth or the fifth. Due to the slow decay of the plucked sound, a near-uniform, continuous drone is clearly audible in the background providing the vocalist (and listeners) with a reference point in the tonal framework. The *tanpura* signal energy is observed to be spread over a very large number of regularly spaced partials throughout the spectrum up to about 10 kHz. This leads to certain frequency bands dominated entirely by *tanpura* partials, thus enhancing its loudness relative to the voice.

C. Percussion (Tabla)

The *tabla* consists of a pair of drums, one large base drum, the *bayan*, and a smaller treble drum, the *dayan*. *Tabla* percussion consists of a variety of strokes, often played in rapid succession, each labeled with a mnemonic. Two broad classes of strokes, in terms of acoustic characteristics, are: 1. tonal strokes that decay slowly and have a near-harmonic spectral structure (thus eliciting a pitch percept) and 2. impulsive strokes that decay rapidly and have a noisy spectral structure. The pitch percept elicited by tonal *tabla* strokes falls within the pitch range of the human singing voice. The acoustic characteristics of various *tabla* strokes were studied from Parag Chordia's database available at http://ccrma.stanford.edu/~pchordia/tablaStrokes/. It was found that while all the impulsive strokes had similar acoustic characteristics, there was a large variability in those of the different tonal strokes. However, the significant harmonic content in the tonal strokes was only found below 1.5 kHz. The local signal-to-interference ratios (SIRs) around the onsets of individual *tabla* strokes can be as low as -10 dB.

III. Audio Data For Experiments

A typical north Indian classical music performance starts at a low tempo where only the voice and *tanpura* are present. The *tabla* strokes, if present, during this time are spaced widely apart in time. As the performance progresses the *tabla* strokes are more closely interspersed and the vocalist sings at a medium tempo. Towards the end of the performance the *tabla* playing becomes very rapid as does the speed of singing.

With this in mind, the data selected for the subsequent studies on preprocessing, classification and segmentation are excerpts from the start, middle and end of recordings of seven different north Indian classical vocal performances spanning 23 minutes in total. Three of the recordings are of female artists and four are of male artists. The recordings are comprised of sounds generated by a single vocalist, a pair of *tablas* and a *tanpura* and, in some cases, a secondary melodic instrument (SMI). The statistics of the vocal and instrumental segments across the entire data set is shown in Table 1. As we can see, the vocal segments comprise nearly 75% of the data.

IV. Pre-processing

We see from the previous section that the accompanying instruments share several temporal and spectral properties with the singing voice. In the vocal regions, the instrumental background has a clear presence and can dominate the voice during softly sung regions. It is of interest therefore to look for methods to enhance the singing voice when present to facilitate the reliable detection of vocal regions.

The *tanpura* spectrum is strongly harmonic and overlaps in range with the voice spectrum, particularly below 5 kHz, ruling out the use of linear filtering. Most north Indian classical music recordings however have initial segments, of duration 4 seconds or more, where only the *tanpura* signal is present. This, along with the relative stationarity of the *tanpura* spectra, directs us towards investigating a spectral subtraction-based *tanpura* suppression scheme.

Spectral subtraction [8] is a well known technique for noise suppression and has been used extensively in speech processing. As its name implies, it involves the subtraction of an average noise power spectrum, estimated during non-speech regions, from the power spectrum of the noisy signal. The enhanced signal is reconstructed from the modified magnitude spectrum and original phase spectrum of the noisy signal. The assumptions made are that the noise is additive and stationary to the degree that its spectrum in the non-speech regions is similar to that during speech. In the present context, the average noise (*tanpura*) magnitude spectrum is estimated over the initial 4 seconds of the audio track. Such a long segment is used to average out the effects of plucking strings tuned to different fundamental frequencies (F0s). This spectrum is then subtracted from all subsequent spectra followed by half-wave rectification and finally signal reconstruction using the overlap-add (OLA) method.

Fig. 1 shows the effect of spectral subtraction based *tanpura* suppression for a typical north Indian classical vocal segment. The figure on the left shows the time domain waveform and narrow-band spectrogram of a segment of music where there is only *tanpura* present from 0 to 5 sec and both *tanpura* and the singer's voice from 5 to 14 sec. The figure on the right shows the time domain waveform and narrowband spectrogram of the reconstructed signal. Spectral subtraction reduces the noise floor, as is evident from the attenuated waveform in the reconstructed signal (from 0 to 5 seconds in the figure on the right). The *tanpura* spectrum is now reduced to a low level of residual noise due to the partial subtraction of its harmonics. We see that the voice harmonics are relatively unaffected.

The *tabla* strokes due to their highly non-stationary characteristics survive the spectral subtraction. Since the major frequency content of the *tabla* lies below about 1500 Hz, the features used for classification are chosen such that they emphasize the difference in the energies of the voice and the *tabla* above 1500 Hz.

V. Frame- Level Classification

For classification we make use of the Gaussian Mixture Model (GMM) classifier. In this section we discuss the need for subdividing vocal and non-vocal classes further, features extracted and the evaluation of the classifier performance on the given data.

A. *Classes*

The objective of the classification stage is to label individual audio frames as vocal (V) or non-vocal (NV). However, within each of these classes the variability in audio characteristics due to the varying nature of the sound production mechanism (for singing) and the different sound sources (for instruments) is considerable. For example, a stop and a vowel, both fall in the vocal class but their audio characteristics are very different. In order to account for this variability we further subdivide the V and NV classes into four sub-classes each.

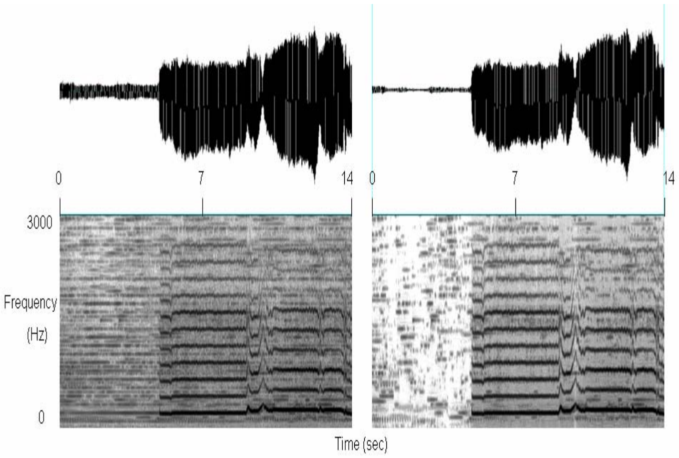

Fig. 1. Time domain waveform (above) and narrowband spectrogram (below) of a segment of a typical north Indian classical recording (left) and the *tanpura* suppressed version (right). The initial segment, from 0 to 5 sec, contains only *tanpura*

The V class is divided into (1) vowels, (2) semivowels and nasals, (3) stops and (4) fricatives. The NV class is divided into (5) *tanpura* only, (6) *tanpura* and SMI, (7) *tanpura* and *tabla*, and (8) *tanpura*, *tabla* and SMI. The number of tokens (feature vectors) available per class after manual annotation of the dataset based on the above 8 classes is 45278, 3807, 1292, 507, 5310, 4472, 1204 and 6996 respectively.

B. *Features*

Based on knowledge of the acoustic characteristics of the underlying classes, seven features are extracted per frame (20 ms) of audio using 40 ms analysis window.

Spectral roll-off is the frequency below which X% of the signal energy is concentrated [9]. 70% was found to provide good separation between tonal *tabla* strokes, *tanpura* and the voice. This feature is defined as

$$\sum_{k=k_0}^{k_{Fr}} \left|X_r[k]\right|^2 = 0.70 \times \sum_{k=k_0}^{k_{Fn}} \left|X_r[k]\right|^2 \qquad (1)$$

where Fr is the roll-off frequency. Fn is the Nyquist frequency. kF, is the spectral bin number whose center is nearest to F Hz. |Xr[k]| is the magnitude spectral value of the kth spectral bin for the rth frame. Lower values of this feature are expected for the frames dominated by the *tabla* and higher values for the frames dominated by *tanpura*. This was especially because the *tabla* has high energy content in the lower frequencies.

Harmonic energy (HE) is sum of the strength of individual harmonics, till 5 kHz, given the predominant pitch. The predominant pitch is computed using the algorithm described in [1]. It is given by

$$HE = \sum_{k=k_{F0}}^{k_{NF0}} \left|X\left[k\right]\right|^2 \qquad (2)$$

where *F0* is the fundamental frequency and *NF0* is the largest multiple of *F0* below 5 kHz. The spectral bin numbers (*k*) corresponding to the closest local maxima, within a 3% neighbourhood, of the expected harmonic location are used in the computation. This feature is expected to have high value during the voiced sounds and lower values in the instrumental regions.

Sub-band energy ratio (ER) is the ratio of the energy of a two frequency bands in spectrum. We have used 2 of these features, one giving the ratio of the energy of a band ranging from 5 kHz to 8 kHz to the energy of a band ranging from 0 kHz to 1.5 kHz and the other giving the ratio of the energy of a band ranging from 2.5 kHz to 5 kHz to the energy of a band ranging from 0 kHz to 1.5 kHz. These are given by

$$ER1 = \frac{\sum_{k=k_{5000}}^{k_{8000}} \left|X_r(k)\right|^2}{\sum_{k=k_0}^{k_{1500}} \left|X_r(k)\right|^2} \; ; \; ER2 = \frac{\sum_{k=k_{2500}}^{k_{5000}} \left|X_r(k)\right|^2}{\sum_{k=k_0}^{k_{1500}} \left|X_r(k)\right|^2} \qquad (3)$$

Lower values of this feature are expected during frames where the voice is present since the concentration of harmonic energy for the voice is below 5 kHz.

Sub-band flux is the spectral flux, as defined in [9], computed over a frequency band ranging from 1.5 kHz to 5 kHz. It is defined as

$$Flux = \sum_{k=k_{1500}}^{k_{5000}} \left(\left| X_r(k) \right| - \left| X_{r-1}(k) \right| \right)^2 \tag{4}$$

It is expected that this feature will generally have a high value during the sung segments, whose spectra show significant inter-frame variation as compared to the more stationary purely instrumental segments.

Audio spectral flatness (ASF) [9] ASF of a spectral band *b* is defined as the ratio of the geometric and the arithmetic means of the power spectrum coefficients within that band. It is given by

$$ASF_r(b) = \frac{\sqrt[(ih(b)-il(b)+1)]{\prod_{k=il(b)}^{ih(b)} |X_r[k]|^2}}{\sum_{k=il(b)}^{ih(b)} |X_r[k]|^2 \Big/ \left(ih(b) - il(b) + 1 \right)} \tag{5}$$

where *ih(b)* and *il(b)* are the highest and lowest frequencies of the spectral band *b*. It is a measure of deviation of the spectral form from that of a flat spectrum. Flat spectra correspond to noise or impulse-like signals. Thus high flatness indicates noisiness. Low flatness values generally indicate the presence of harmonic components. The band chosen for classification was 1.5 kHz – 3 kHz. Voice has a harmonic structure until 5 kHz and hence gives a high value for this feature. As mentioned before, the dominant harmonics of the tonal *tabla* strokes are present below 1.5 kHz. Hence this feature was computed over the band ranging from 1.5 – 3 kHz.

Average sub-band energy (SE) is the average energy of the spectral sub-band ranging from 1.5 kHz to 5 kHz per bin.

The Shannon Mutual Information values [10] for each of the above features are estimated as 0.2747 (roll-off), 0.6016 (HE), 0.2437 (ER1), 0.2402 (ER2), 0.3939 (Flux), 0.4680 (ASF) and 0.4484 (SE) respectively. Also, the correlation between features was estimated to be low.

C. *Evaluation*

We perform 10-fold cross validation to evaluate the overall performance of the classifier. A variable number of mixtures per sub-class are used to train the classifier. The number of mixtures per sub-class finally used was arrived at by trials over a representative set of mixtures whose upper limit was governed by the number of tokens per sub-class available. Here we used 10, 6, 2, 1, 8, 7, 2 and 10 mixtures for sub-classes (1) to (8) respectively.

Table 2 shows the overall confusion matrix for the resulting vocal and non-vocal classes is computed by grouping the results for individual sub-classes over all 10 folds.

Table 2. Overall confusion matrix for frame-level V/NVclassification

	V (Classified)	NV (Classified)
V (Ground truth)	73.70 %	26.30 %
NV (Ground truth)	27.47 %	72.54 %

VI. Audio Segmentation And Segment Classification

Analysis windows, used in short-term audio analysis, should be as short as possible so that the signal within the window is almost stationary. However, short-term classification, as described in the previous section, is not entirely reliable since the information within a frame is limited. The sensation of a sound texture or long-term nature of sound occurs due to the combined effect of temporal and spectral characteristics over a longer duration.

In order to capture this, we take the approach of segmenting the audio record into homogenous regions and then using the frame-level classifier output over these regions to classify the segment as vocal or non-vocal. The segment boundary detection is achieved by locating abrupt changes in suitably chosen short-term features.

A. *Implementation*

The two features used for segmentation are sub-band flux and sub-band energy ratio (ER) extracted using a fixed window length of 30 ms and a frame rate of 100 frames/sec. These are defined as in Section *V.B.* The bands for the energy ratio computation now range from 5 kHz to 9 kHz and from 1.5 kHz to 5 kHz. For boundary marking using SE, we use an upper and lower fixed threshold. Empirical values of these thresholds that were found to give good results are 0.3 and 0.2. Boundaries are marked at frames where the ER crosses a value of 0.3 with positive slope or crosses 0.2 with a negative slope. For boundary marking using Flux, again we use two fixed thresholds. Empirical values of these thresholds that were found to give good results are 40 (for upper) and 20 (for lower) dB. Boundaries are marked at frames where the sub-band flux crosses a value of 40 dB with positive slope or crosses 20 dB with a negative slope.

The final set of boundaries is the union of the boundaries detected using each of the above features individually. Further, this set is pruned by examining the boundaries in a left-to-right manner and applying the following rule. If two boundaries are found to occur within 100 ms of each other then the later boundary is discarded. This rule is based on the observation that the duration of a sung phrase or a purely instrumental inter-phrase break almost always exceeds 200 ms. The marking of these boundaries within short intervals of each other was caused by either successive strong *tabla* stroke onsets or pitch variations in the singing voice.

B. *Evaluation of Segmentation Method*

Here, the accurate marking of sung phrase onsets and offsets is of more importance than the reduction of boundaries marked within a homogenous segment (false alarms). We have observed that the duration of a sung phrase rarely falls below 1 sec. A phrase onset or offset is said to be correctly determined when the nearest marked boundary lies within a neighbourhood of 150 ms. Using this criterion we found that 78.9 % of actual phrase boundaries have been correctly identified.

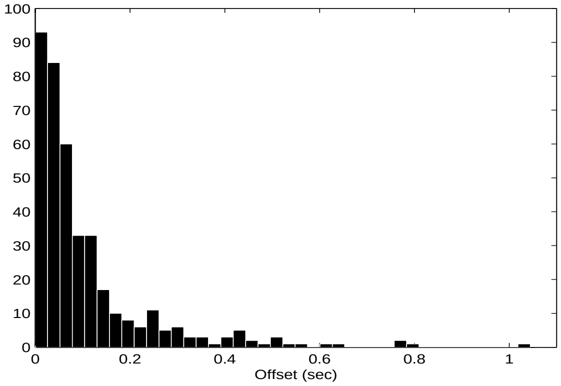

Fig. 2. Histogram of boundary location offsets

Fig. 2 shows a histogram of the offsets from actual boundaries, obtained from annotated data, occurring during boundary marking. We can see that the majority of the boundaries are marked within 200 ms of the manually marked boundaries. There are also few very high offsets (around 1.1 sec). These occur during audio segments where the Signal-to-Interference Ratio (SIR) is very low i.e. where the voice is very soft. This happens sometimes during the end of sung phrases where the voice gradually trails off.

C. *Segment Classification Results*

The first method of grouping frame-level classification results over segments used here is majority voting (MV) i.e. if the majority of frames of a portion are classified as vocal the portion is classified as vocal, and vice versa. The second method used involves grouping the frame-level log likelihoods (LLH) for each sub-class over the segment. A segment is classified as vocal if the maximum overall likelihood of all sub-classes occurs for a vocal sub-class.

For evaluation of segment classification, the classifier was trained using the entire data set. Then all single audio excerpts were individually input as testing data to the classifier. Overall confusion matrices were computed before grouping, after grouping using majority vote and using log likelihoods. These are shown in Table 3.

The boundary marker used detects several extra boundaries in addition to the sung phrase boundaries. To get a perspective on how much improvement in segment classification is possible by improving the performance of the boundary marker, Table 4 shows the confusion matrices for grouping using majority vote and using log likelihoods using the boundary markers extracted from the manual annotation of individual excerpts (ideal boundaries).

Table 3. Overall confusion matrices for segment classification using automatically detected boundaries

		Before grouping		After grouping	
		V	NV	V	NV
MV	V	79.25 %	20.75 %	84.75 %	15.25 %
	NV	25.20 %	74.80 %	21.31 %	78.70 %
LLH	V	79.25 %	20.75 %	82.87 %	17.13 %
	NV	25.20 %	74.80 %	20.44 %	79.56 %

Table 4. Overall confusion matrices for segment classification using manually detected boundaries

		After grouping	
		V	NV
MV	V	92.04 %	7.96 %
	NV	24.24 %	75.76 %
LLH	V	91.12 %	8.88 %
	NV	22.98 %	77.02 %

From Table 3 it can be seen that the accuracy of the system improves when frame-level decisions are grouped over segments. It can also be seen that both methods of grouping studied provide similar results. From Table 4, we note that an improvement in boundary marking will result in still higher classification accuracies.

VII. Summary

The problem of accurate detection of the presence of the singing voice is studied in the context of north Indian classical vocal performances. Based on the signal characteristics of the voice and the *tanpura*, a pre-processing method for *tanpura* suppression was proposed. A GMM classifier using appropriately selected features was evaluated over a representative dataset. Further improvement in classification accuracies was achieved by grouping frame-level decisions over segments, which were automatically detected using the proposed boundary marking algorithm.

We are currently addressing the improvement of the accuracy of boundary detection by investigating new features related to the time trajectory of the tonal components.

References

[1] A. Bapat, V. Rao, and P. Rao, “Melodic contour extaction for Indian classical vocal music,” in *Proc. International Workshop on Artificial Intelligence and Music (IJCAI-07)*, Hyderabad, India, 2007, pp.13-24.

[2] T. Zhang, “System and method for automatic singer identification,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 33-36, 2003.

[3] Y. Kim, and B. Whitman, “Singer identification in popular music using voice coding features,” in *Proc. 5th International Conf. on Music Information Retrieval*, Barcelona, Oct. 10-14, 2004.

[4] A. Berenzweig, D. Ellis, and S. Lawrence, “Using voice segments to improve artist classification of music,” in *Proc. AES 22nd International Conference*, 2002.

[5] N. Maddage, C. Xu, and Y. Wang, “An svm-based classification approach to musical audio,” in *Proc. 4th International Conf. on Music Information Retrieval*, Washington D.C.., Oct. 26-30, 2003.

[6] Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monoaural recordings,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475-1487.

[7] T. New, A. Shenoy, and Y. Wang, “Singing voice detection in popular music,” *Proc. 12th annual ACM International Conference*, New York, Oct. 10-16, 2004.

[8] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Audio, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[9] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” CUIDADO I.S.T. Project Report 2004.

[10] R. Battiti, “Using mutual information for selecting features in a supervised neural net learning,” *IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537-550 , 1994.