Singing Voice Detection using Modulation Frequency Features

Maria Markaki[1,2], *Andre Holzapfel*[1,2], *Yannis Stylianou*[1,2]

[1]Computer Science Department, University of Crete, Greece
[2]Institute of Computer Science, FORTH, Greece
{mmarkaki, hannover, yannis}@csd.uoc.gr

Abstract

In this paper, a feature set derived from modulation spectra is applied to the task of detecting singing voice in historical and recent recordings of Greek Rembetiko. A generalization of SVD to tensors, Higher Order SVD (HOSVD), is applied to reduce the dimensions of the feature vectors. Projection onto the "significant" principal axes of the acoustic and modulation frequency subspaces, results in a compact feature set, which is evaluated using an SVM classifier on a set of hand labeled musical mixtures. Fusion of the proposed features with MFCCs and delta coefficients reduces the optimal detection cost from 11.11% to 9.01%.

Index Terms: audio classification, modulation spectrum, singing voice activity detection.

1. Introduction

Determining the parts of a musical piece in which a melody is sung is referred to as singing voice detection [1, 2]. Being able to locate such parts is of interest in applications like the extraction of small characteristic snapshots from a piece of music or for the recognition of the singer in a bigger collection of musical pieces. In [2] a simple threshold is derived from the harmonic structure of the magnitude of the Fourier transform. In the more recent work by Rocamora et.al. [1] an overview of different features for the detection of singing voice is given, and it is summed up that MFCC with delta coefficients appear the superior feature set for this task.

The task of singing voice detection has a close relation to the task of speech/non-speech segmentation. Here also, MFCC features have been successfully applied [3, 4]. In [5] features derived from modulation spectra have been shown to improve in this task compared to MFCC. Hence in this work we evaluate this type of feature set in a singing voice detection task.

A modulation spectrum based description of a signal captures fast and slower time-varying quantities such as pitch, phonetic and syllabic rates of speech, tempo of music, etc [6]. Still, the use of modulation spectral features in pattern classication is prevented by their large dimensionality. In this paper, a generalization of SVD to tensors (Higher Order SVD [7]) reduces the dimensionality of the features. This technique has been applied in auditory-based features with multiple scales of time and spectral resolution [8]. Joint acoustic and modulation frequencies are projected on the retained singular vectors in each subspace to obtain the multilinear principal components (PCs) of the sound samples. Next we examine the relevance to the target class of the largest PCs in the acoustic frequency and the modulation frequency subspace using a mutual information based criterion.

This compressed modulation frequency representation is evaluated using a hand labelled data set previously used for vocal frame selection and singer recognition in [9]. The performance of an SVM classifier is presented, giving emphasis on a detailed illustration of its behaviour regarding measures like the Detection Error Trade-off curves [10], previously more common in Speech/non-Speech related publications than in Music Information Retrieval.

The organization of the paper is as follows: Section 2 briey reviews the modulation frequency analysis framework. The multilinear dimensionality reduction and the mutual information estimation method are presented in Section 3. In Section 4 we describe the experimental setup, the database and the results. Finally in Section 5 we present our conclusions.

2. Modulation Frequency Analysis

The most common modulation frequency analysis framework [6] for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) $X_k(m)$

$$\begin{aligned} X_k(m) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn}, \qquad (1) \\ k &= 0,\ldots,K-1, \end{aligned}$$

where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window. Subband envelope detection - defined as the magnitude $|X_k(m)|$ or square magnitude of the subband - and their frequency analysis with Fourier transform are performed next:

$$\begin{aligned} X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_k(m)|W_I^{im}, \qquad (2) \\ i &= 0,\ldots,I-1, \end{aligned}$$

where $g(m)$ is the modulation frequency analysis window; k and i are referred to as the "Fourier" (or acoustic) and "modulation" frequency, respectively. Tapered windows $h(n)$ and $g(m)$ are used to reduce the sidelobes of both frequency estimates.

A modulation spectrogram representation then, displays modulation spectral energy $|X_l(k,i)|$ in the joint acoustic/modulation frequency plane. Length of the analysis window $h(n)$ controls the trade-off between resolutions in the acoustic and modulation frequency axes. The degree of overlap between successive windows sets the upper limit of the subband sampling rate during the modulation transform.

3. Description of the method

3.1. Multilinear Analysis of Modulation Frequency Features

Every signal segment in the training database is represented in the acoustic-modulation frequency space as a two-dimensional

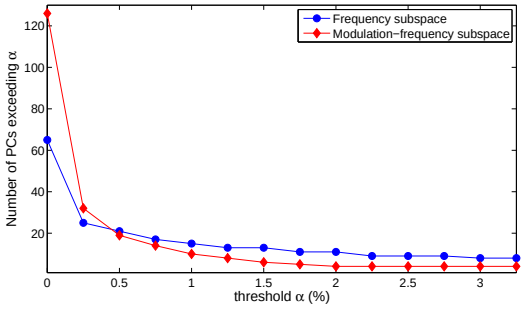

Figure 1: *Total number of retained PCs in each subspace as a function of threshold on contribution percentage. The vertical axis indicates the number of PCs in each subspace that have contribution (eq.5) greater than the threshold*

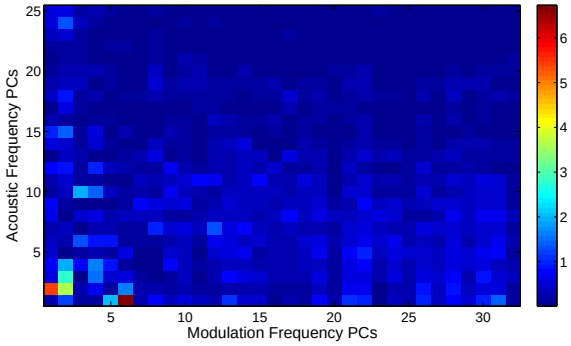

Figure 2: *MI between projections of features and the class variable, divided by the median value of MI between pairs of projections.*

matrix. By stacking all training matrices we obtain a data tensor. A generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [7] enables the decomposition of a tensor D to its mode$-n$ singular vectors:

$$D = S \times_1 U_{frequency} \times_2 U_{mod_freq} \times_3 U_{samples} \quad (3)$$

where $U_{frequency}$, and U_{mod_freq} are the orthonormal ordered matrices of the corresponding subspaces of acoustic and modulation frequencies; these contain subspace singular vectors, obtained by unfolding D along its corresponding modes. Samples subspace matrix, $U_{samples}$, is ignored. Tensor S is the core tensor with the same dimensions as D. $S \times_n U$ where $n = 1, 2, 3$ denotes the $n-$ mode product of tensor $S \in R^{I_1 \times I_2 \times I_3}$ by the matrix $U \in R^{J_n \times I_n}$. For $n = 2$ for example, it is an $(I_1 \times J_2 \times I_3)$ tensor given by

$$(S \times_2 U)_{i_1 j_2 i_3} = \sum_{i_2} s_{i_1 i_2 i_3} u_{j_2 i_2}. \quad (4)$$

Each singular matrix can be truncated then by setting a predetermined threshold so as to retain only the desired number of principal axes in each mode. The contribution of the j^{th} principal component (PC) of subspace S_i whose corresponding eigenvalue is $\lambda_{i,j}$, is defined as:

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{j=1}^{N_i} \lambda_{i,j}} \quad (5)$$

where N_i is the dimension of S_i - 65 for acoustic frequency and 126 for modulation frequency. Figure 1 presents the number of PCs in these two subspaces as a function of $\alpha_{i,j}$.

Joint acoustic and modulation frequencies $B_{mod}[f, t]$ extracted from new sound samples are first mean subtracted (mean values estimated from the whole training set) before they are projected on the truncated orthonormal axes of interest, U'_{freq} and U'_{mod_freq}

$$Z = B \times_1 {U'_{freq}}^T \times_2 {U'_{mod_freq}}^T \quad (6)$$

The resulting matrix Z whose dimension is equal to the product of retained singular vectors in each mode contains thus the multilinear PCs of a sound sample.

Next, we detect the near-optimal projections (principal components) of features among those contributing more than 0.25% based on mutual information [13, 11]. That is, we examine the relevance to the target class of the first 25 PCs in the acoustic frequency subspace and the first 32 PCs in the modulation frequency subspace.

3.2. Mutual Information Estimation

The mutual information (MI) $I(x_i; x_j)$ between two random variables x_i and x_j is defined in terms of their joint probability density function (pdf) $P_{ij}(x_i, x_j)$ and the marginal pdf's $P_i(x_i)$, $P_j(x_j)$:

$$I(x_i; x_j) = \sum_{x_i, x_j} P_{ij}(x_i, x_j) \log_2 \left(\frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \right) \quad (7)$$

It coincides with the Kullback Leibler divergence, a measure of distance, between $P_{ij}(x_i, x_j)$ and the product of $P_i(x_i)$, $P_j(x_j)$. Mutual information then is a natural measure of the inter-dependence between those variables.

Estimating MI from a finite sample requires regularization of $P_{ij}(x_i, x_j)$. The simplest regularization is to define b discrete bins along each axis. We make an adaptive quantization (variable bin length) so that the bins are equally populated and the coordinate invariance of the MI is preserved [11]. The precision of features quantization also affects the sample size dependence of MI estimates [14]. We have set $b^* = 8$ according to a procedure described in [11]: when data are shuffled, mutual information should be near zero for $b < b^*$ while it increases for $b > b^*$.

We estimate $I(x_i; c)$, the MI between each of the projections of modulation spectra on the first 25×32 PCs and the class variable (singing voice vs instrumental music) for the training set. We also address the "similarity" between features, estimating the mutual information $I(x_i; x_j)$ between pairs of features x_i and x_j [11]. A simple measure of the redundancy of each feature x_i then, is the median (or mean) value of its "similarity" $I(x_i; x_j)$ to the other features, $RS(x_i)$. The fraction $I(x_i; c)/RS(x_i)$ reflects the optimality of feature x_i for classification: its relevance to the target class scaled down by its redundancy. Figure 2 shows that the subspace spanned by the first ~ 10 acoustic frequency PCs and the first ~ 6 modulation frequency PCs includes the "optimal" projections. This subspace roughly corresponds to the PCs with eigenvalues contributing more than 1.5% (see eq.(5) and Figure 1).

3.3. System evaluation

Classification of segments was performed using support vector machines.We have used SVMlight [12] with a Radial-Basis-Functions kernel.

We evaluate system performance on the test set using the detection error trade-off curve (DET) between false rejection rate (or speech miss probability P_{miss}) and false acceptance rate (or false alarm probability P_{false}) [10] . Since prior probability of singing voice class in our test data set is $P_{target} = 55.43\%$, if the costs of miss and false alarm probabilities are considered equally important, the minimum value of the detection cost function, DCF_{opt}, according to [10], is:

$$DCF_{opt} = \min\left(\frac{P_{miss}.P_{target}+P_{false}.(1-P_{target})}{2}\right). \quad (8)$$

4. Experiments

4.1. Data Collection

The data set used to evaluate the system for voice activity detection contains historical and recent recordings of Greek Rembetiko[1] music. It was used in [9] for singing voice activity detection in the framework of a singer recognition system. It consists of 84 songs from 21 singers. A test set of 21 songs, one from each singer, has been separated, leaving a total of 63 songs for the development of the system.
All 84 songs have been hand labelled with the following labels:

- INSTR : instrumental sounds without any voice
- VOICE : voice of target singer without second voice
- MIXED : voice of target singer with second voice
- OTHER : interjections

The focus will lie on the distinction between instrumental frames without any voice and frames with some kind of voice activity. Each file has been partitioned into 1000 ms segments for long-term feature analysis, producing 12500 samples for training (and validation), and 3763 samples for testing.

4.2. Feature Extraction and Classification

The modulation spectrogram has been calculated using Modulation Toolbox [16]. For every 1000 ms block modulation spectrum features were generated using a 128 point spectrogram with a Gaussian window. The envelope in each subband was detected by a magnitude square operator. To reduce the interference of large dc components of the subband envelope, the mean was subtracted before modulation frequency estimation. One uniform modulation frequency vector was produced in each one of the 65 subbands. Due to a window shift of 32 samples, each modulation frequency vector consists of 126 elements up to 250 Hz. Joint acoustic and modulation frequencies are projected on the truncated orthonormal axes U'_{freq}, and U'_{mod_freq} according to eq. (6). All features were normalized by their corresponding standard deviation estimated from the entire training set to reduce their dynamic range before classification with SVMs.

4.3. Results on the Validation Set

Table 1 presents the minimum detection cost function (DCF_{opt}) and the false rejection rate for low false acceptance rate on the validation set when retaining PCs with contributions greater than 0.25%, 0.5%, ... up to 3.25% in 0.25% steps (see Figure 1). The dimensionality of the reduced features progressively decreases from 800 to 15 features; up to ~ 80 features, classification error decreases due to improved SVM generalization. However with less than $[13 \times 6]$ PCs, the performance degrades especially in terms of false rejection probability at low false alarm rates. This can be attributed to the loss of highly informative PCs in each subspace, as depicted in Figure 2. Feature selection according to [13] did not yield any advantage over the first $[13 \times 6]$ PCs(results not shown). Probably the reason is that the first $[13 \times 6]$ PCs include the most informative PCs in both subspaces. Since SVM exhibit a good generalization performance for up to ~ 100 features, there is no obvious advantage in reducing dimensionality by feature selection [12].

[1] http://www.rebetiko.gr/en/history.php

Table 1: *Classification results on validation set*

System	DCF	$FR@FA = 2\%$	$FR@FA = 1\%$
$[25 \times 32]$	35.21 %	92.33%	95.78%
$[21 \times 19]$	23.09%	86.48%	89.79%
$[17 \times 14]$	20.84%	82.02%	87.1%
$[15 \times 10]$	19.32%	73.39%	82.6%
$[13 \times 8]$	20.01%	70.04%	80.01%
$[13 \times 6]$	19.98%	67.11%	75.46%
$[11 \times 5]$	20.73%	71.57%	78.81%
$[11 \times 4]$	22.08%	71.76%	79.58%
$[9 \times 4]$	24.1%	73.92%	82.02%
$[8 \times 4]$	24.21%	74.98%	84.32%
$[7 \times 3]$	26.12%	79.53%	85.09%
$[6 \times 3]$	27.11%	80.3%	87.3%
$[5 \times 3]$	28.01%	85.47%	90.6%

4.4. Combining Modulation and Cepstral Features

A comparative study on audio descriptors for singing voice detection [1], concluded that the most appropriate feature set were the median and standard deviation of MFCCs and their delta coefficients, estimated over 1 second segments. Moreover the authors reported that combination of different descriptors did not improve classification performance. We test here whether the modulation features could provide non-redundant information to that encoded by MFCCs, for this particular task.

We derive 13 coefficients from 40 mel scale frequency bands in overlapping frames of 25 ms with 10 ms hop size. We also apply equal loudness pre-emphasis and cubic-root amplitude compression according to [15], implemented using [17]. Based on the results of the previous experiments, we combine the modulation features projected onto the first $[13 \times 6$ PCs ($78-$dimensional) with the $52-$dimensional MFCC features. We simply concatenate the two feature vectors prior to classification with SVMs. All features were normalized by their corresponding standard deviation estimated from the entire training set. The respective DET curves are shown in Figure 5. Modulation features give an improvement over different decision thresholds, particularly in the low false alarm region. Table 2 presents the optimal values of DCF, P_{miss} and P_{false} for the systems tested as well as the false rejection at low false alarm rates. Table 3 presents the accuracy of vocal frames selection based on the decision threshold corresponding to a false alarm rate of 0.5%.

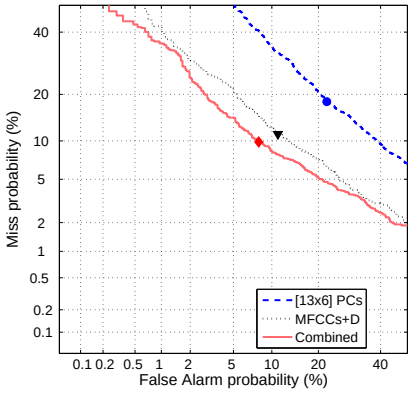

Figure 3: *Detection Error Trade-off curves and minimum detection cost function (markers) for the test-set (see Section 4.1) when retaining PCs with contribution greater than* 1.5% *(circle), using MFCCs + delta coefficients (△) and the combined set of features (□)*

Table 2: *Classification performance on test set*

	$[13 \times 6]$ PCs	MFCCs + D	Combined
DCF	19.83%	11.11%	9.01%
P_{miss}	18.01%	11.17%	9.83%
P_{false}	22.11 %	11.03%	7.99%
EER	21.74 %	13.73%	12.06%
$FR@FA = 0.5\%$	81.26 %	51.92%	43.48%
$FR@FA = 1\%$	75.11 %	40.89%	35.95%
$FR@FA = 2\%$	67.01 %	30.49%	25.79%

5. Discussion and Conclusions

This paper presented a novel feature set for the detection of singing voice in old and new musical recordings.In a speech - nonspeech discrimination task [5], these "reduced" modulation features exhibited comparable classification performance to that of "perceptual" MFCCs [15]. In the case of singing voice discrimination from other harmonic musical instruments, modulation features were inferior to MFCCs and their delta coefficients, the feature set most suitable for this task according to a recent comparative study [1]. Still, their combination improved classification results especially in the low false alarm region - which is important in singer recognition applications. The results support the hypothesis that modulation features provide non-redundant information to that encoded by MFCCs.

6. References

[1] Rocamora, M. & Herrera, P., "Comparing audio descriptors for singing voice detection in music audio files", Proc. 11th Brazilian Symposium on Computer Music, San Pablo, Brazil, 2007.

[2] Tin Lay Nwe, Arun Shenoy & Ye Wang, "Singing voice detection in popular music", MULTIMEDIA '04: Proc. 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004.

[3] Lu, L., Zhang, H.J. & Li, S., "Content-based audio classification and segmentation by using support vector machines", Multimedia Systems 8: 482-492, 2003.

[4] Aronowitz, H., "Segmental modeling for audio segmentation", Proc. ICASSP 2007, Hawaii, USA, 2007.

[5] Markaki, M., & Stylianou, Y., "Discrimination of speech from nonspeech in broadcast news based on modulation frequency features", Proc. ISCA, Speech Analysis and Processing for Knowledge Discovery, June 2008 (accepted).

Table 3: *Classification accuracies on test set*

method	accuracy	correct frames	false accepted
$[13 \times 6]$ PCs	97.99%	391	8
MFCCs + D	99.17%	1003	8
Combined	99.29%	1179	8

[6] Atlas, L. & Shamma S.A., "Joint Acoustic and Modulation Frequency", EURASIP Journal on Applied Signal Processing, 7:668–675, 2003.

[7] De Lathauwer, L., De Moor, B. and Vandewalle, J., "A multilinear singular value decomposition", SIAM J. Matrix Anal. Appl., vol. 21, pp. 1253–1278, 2000.

[8] Mesgarani, N., Slaney, M. & Shamma S.A., "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. Audio, Speech and Language Proc., 14:920–930, 2006.

[9] Holzapfel A. & Stylianou Y., "Singer Identification in Rembetiko Music", Proc. of SMC 2007, Conference on Sound and Music Computing, Lefkada, Greece, 2007.

[10] The NIST Year 2004 Speaker Recognition Evaluation Plan, "http://www.nist.gov/speech/tests/spk/2004/"

[11] Slonim, N., Atwal, G.S. , Tkacik, G. & Bialek, W. "Estimating mutual information and multi-information in large networks", *"http://arxiv.org/abs/cs.IT/0502017"*, 2005.

[12] Joachims, T., "Making large-scale SVM Learning Practical" *in Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, eds., MIT-Press, 1999.

[13] Peng, H., Long, F. & Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", IEEE Trans. Pattern Analysis and Machine Intelligence, vol 27, 8:1226–1238, 2005.

[14] Cover, T.M. & Thomas, J.A., Elements of Information Theory, John Wiley and Sons, New York, 1991.

[15] Hermansky, H., Hanson, B. & Wakita, H., "Perceptually based linear predictive analysis of speech", Proc. ICASSP 1985, pp. 509-512, 1985.

[16] Modulation Toolbox : "http://www.ee.washington.edu/ research/isdl/projects/modulationtoolbox"

[17] "http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat"