Unsupervised Singing Voice Detection Using Dictionary Learning

Aggelos Pikrakis
Dept. of Informatics
University of Piraeus, Greece
email: pikrakis@unipi.gr

Yannis Kopsinis
Libra MLI
Edinburgh, UK
email: kopsinis@ieee.org

Nadine Kroher and José-Miguel Díaz-Báñez
Dept. of Applied Mathematics II
University of Seville, Spain
email: {nkroher, dbanez}@us.es

***Abstract*—This paper presents an unsupervised approach to vocal detection in music recordings based on dictionary learning. At a first stage, the recording to be segmented is treated as training data and the K-SVD algorithm is used to learn a dictionary which sparsely represents a short-term feature sequence that has been extracted from the recording. Subsequently, the vectors of the feature sequence are reconstructed based on the learned dictionary and the probability of appearance of the dictionary atoms is estimated. The obtained probability serves to compute the value of a weight function for each frame of the recording. The histogram of this function is then used to estimate a binarization threshold that segments the recording into vocal and non-vocal segments. The performance of the proposed unsupervised method, when evaluated on two datasets of accompanied singing, presents comparable performance to supervised techniques.**

I. Introduction

Singing voice detection or vocal detection, abbreviated SVoD in this paper, refers to the task of detecting automatically the segments of a music recording where the singing voice is present. The singing voice is the central element in various music genres and, consequently, a reliable SVoD method can be a key component for a number of tasks in the field of Music Information Retrieval (MIR), including singer identification, singing melody transcription, lyrics alignment and structural segmentation, to name but a few.

Prior research has mainly approached SVoD as a binary, supervised classification task ([1], [2], [3], [4], [5], [6], [7], [8], [9]). In these methods, a model is first trained on features extracted from an annotated corpus and the model is then applied to classify each frame of an unknown recording to the voiced or unvoiced class. To this end, the discriminative capability of various features has been investigated, including spectrally derived low-level representations (MFCCs, LPCCs, PLPs, LFPCs) and more elaborate features, like the pitch contour, vibrato, tremolo and signal envelope. On the other hand, unsupervised approaches have been rare. Specifically, the work in [10] presents a method for discriminating among vocals and static accompaniment based on the fluctuation of harmonic partials and the method in [11] estimates vocal sections by tracking the presence of vibrato in the harmonics of the spectral representation of the signal.

In this paper, we present an unsupervised SVoD method based on Dictionary Learning (DL). DL is directly related to the sparse representation/sparse coding tasks, the objective of which is to choose a few elementary signals/vectors, called atoms, drawn from a pre-specified set, referred to as dictionary. This dictionary provides a data-aware compact representation of a data set in terms of the atoms. The popularity of DL methods in diverse application contexts has also been reflected, within a certain extent, to tasks related to MIR and signal processing for music. For example, DL methods have been employed to yield sparse representations in the context of source separation [12], genre classification [13], [14], [15], music annotation/retrieval [16], multi-pitch analysis [17] and music transcription [18], [19], [20].

The novelty of our approach lies in the fact that the short-term feature sequence representing the signal to be segmented is treated as unlabelled training data which are used to learn, in an unsupervised mode, a small dictionary using the well known K-Singular Value Decomposition (K-SVD) algorithm [21]. The learned dictionary is then employed to reconstruct the original short-term representation. The resulting sparse reconstruction matrix reveals the probability of appearance of each dictionary atom over the entire reconstructed recording and therefore, the atom's respective information content. Every atom combination that reconstructs a short-term feature vector is then mapped to a positive value based on a scheme that weights the information content of the participating atoms with the normalized intensity of their contribution in the reconstruction of the frame. The histogram of the resulting function over the short-term frames exhibits peaks and valleys and is used to compute a global threshold based on the Otsu method from the field of image analysis [22]. This threshold is subsequently used to binarize the feature sequence. Therefore, the proposed approach circumvents the need for a training stage on previously annotated data. Our experiments on two datasets of accompanied singing have shown that despite its unsupervised nature, the proposed method is competitive to supervised approaches.

The paper is structured as follows: Section II presents the DL methodology, Section III describes the segmentation stage and Section IV presents the experimental setup and the respective performance evaluation. Finally, conclusions are drawn in Section V.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

II. Dictionary Learning

Let D be a dictionary matrix comprising r dictionary atoms, $\mathbf{d}_1, \cdots, \mathbf{d}_r$ as columns. Dictionary learning refers to the estimation of D via an optimization task which is often formulated as follows [21], [23], [24]:

$$\min_{D,B} \; || \, X - DB \, ||_{\mathrm{F}}^2, \text{ s.t. } || \, \mathbf{b}_j \, ||_0 \leq k, \; j = 1 \cdots N, \quad (1)$$

where X contains the training vectors ($\mathbf{x}_i, \; i = 1 \ldots N$) as columns, $|| \cdot ||_{\mathrm{F}}$ is the Frobenious norm, $|| \cdot ||_0$ is the ℓ_0 pseudo-norm counting the number of nonzero components of the unknown vector $\mathbf{b}_j$ and k is the number of atoms, which are linearly combined to represent each training vector. Moreover, a constraint on the dictionary norm is necessary in order to avoid degenerate solutions, with the unit-norm request for each atom being the most popular. In this paper, X comprises a set of l-dimensional feature vectors that have been extracted from the music recording using a short-term feature extraction scheme. We experimented with the bark band representation of the spectrogram of the recording and the well known Mel Frequency Ceptrum Coefficients (MFCCs). Details are given in Section IV.

The dictionary is usually trained in an iterative fashion, alternating two learning stages until convergence. In the first stage, the dictionary D is fixed to its latest estimate and B is estimated, column by column, via a series of sparse coding tasks, i.e.,

$$\min_{\mathbf{b}_j} || \, \mathbf{x}_j - D\mathbf{b}_j \, ||^2 \text{ s.t. } || \, \mathbf{b}_j \, ||_0 \leq k, \; j = 1 \cdots N \quad (2)$$

In the second stage, the dictionary is updated, whereas on the same time, either the full matrix B, or its zero entries only, are kept fixed, depending on the specific DL method. One of the most popular and well-performing DL methods, the K-Singular Value Decomposition (K-SVD) [21], updates the dictionary atom by atom via a series of rank-1 approximations performed via truncated SVD.

The power of dictionary learning lies in the fact that *all* the training data vectors are "forced" to be represented with *only a few*, in particular k, dictionary atoms. As a result, the dictionary atoms are rendered highly informative, grasping the "essence" of the available training examples set. After the learning phase, any unseen vector, $\mathbf{a}$, which shares similar structure and/or characteristics with all or a subset of the training vectors, can be sparsely represented as well. In particular, the k-sparse representation of $\mathbf{a}$ is computed via

$$\hat{\mathbf{b}} := \min_{\mathbf{b}} || \, \mathbf{a} - D\mathbf{b} \, || \text{ s.t. } || \, \mathbf{b} \, ||_0 \leq k, \quad (3)$$

The certain atoms used for its representation carry important information regarding its intrinsic characteristics.

III. Singing voice detection

Having assumed a feature sequence of N, l-dimensional feature vectors at the output of a feature extraction stage, a dictionary, D, of r atoms and a full matrix, B, of k-sparse representations (columns), we define a new matrix, F, such that $F_{ij} = 1$ if $B_{ij} \neq 0$. In other words, the j-th column of F has ones on the rows that correspond to the k atoms that reconstruct the j-th frame. Matrix F can be readily used to compute the frequency of appearance of the i-th dictionary atom after the reconstruction procedure has been completed. In the sequel, we will treat this frequency as the probability of appearance, p_i, of the i-th dictionary atom, $\mathbf{d}_i, i = 1, \ldots, r$, computed as

$$p_i = \frac{\sum_{j=1}^{N} F(i,j)}{\sum_{i=1}^{r} \sum_{j=1}^{N} F(i,j)} \quad (4)$$

It follows that the information content, H_i, of the i-th dictionary atom, is

$$H_i = log_2 \frac{1}{p_i} \quad (5)$$

We then compute a function value, $s(j), j = 1, \ldots, M$, for each feature vector, by weighting the information content of each one of the k atoms that participate in the reconstruction of the vector with a weight that represents the intensity of the atom's contribution:

$$s(j) = \frac{\sum_{i=1}^{r} F(i,j) H_i \mid B(i,j) \mid}{\sum_{i=1}^{r} \mid B(i,j) \mid} \quad (6)$$

where $\mid \, . \, \mid$ stands for the absolute value. In this equation, $\frac{|B(i,j)|}{\sum_{i=1}^{r} |B(i,j)|}$ is the weight of the i-th atom during the reconstruction of the j-th vector.

The values of s are then smoothed with a median filter (1s long) and their histogram is computed using a predefined number of bins (256 bins in our study). An example of the histogram of the resulting smoothed sequence is shown in Figure 1.

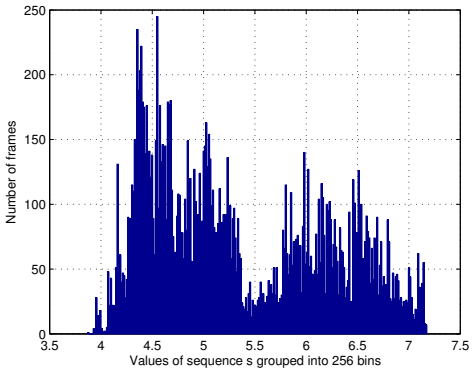


Fig. 1: Histogram of smoothed sequence, s, for the first music recording of the "Cante-100" collection [25], using a 3-sparse representation and a dictionary of 64 atoms to reconstruct the bark band representation of the spectrogram of the recording.

It can be seen that although the histogram exhibits several peaks and valleys, an area of low values can be observed around value 5.5. This area separates the histogram in two parts.

We therefore make the assumption that the left part of such a histogram corresponds to frames from the background

(accompaniment) of the music recording and the right part to the frames of the singing voice. To validate our assumption, we use the well known Otsu method [22], that was originally proposed in the area of image binarization, to compute a single threshold, T_h, that serves to binarize each frame of the feature sequence, i.e., the i-th frame is considered to be voiced, if $s(i) > T_h$. We therefore distinguish the short-term frames into two classes based on the weighted information content that they carry, as this is reflected by the probability of excitation and corresponding reconstruction magnitude of the respective dictionary atoms. For the example of Figure 1, the Otsu threshold is 5.505 and the values of Precision, Recall and F-measure for the class of the singing voice are 96.07%, 94.32% and 95.19%, respectively.

A more complex histogram can be seen in Figure 2. In this case, the Otsu threshold is 5.737 and the values of Precision, Recall and F-measure for the class of the singing voice are 93.32%, 77.71% and 84.83%, respectively, i.e., a performance drop is observed due to a lower recall value, which is, in turn, due to the fact that the computed Otsu threshold is higher than desired because the histogram is more complicated with respect to the observed number of peaks (modalities).

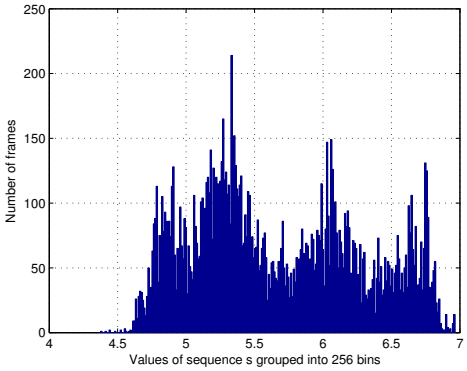


Fig. 2: A more complex histogram of a smoothed sequence, s, for the 51st recording of the "Cante-100" collection [25], using a 3-sparse representation and a dictionary of 64 atoms to reconstruct the bark band representation of the spectrogram of the recording.

Note that, the efficiency of our method relies on the fact that, although the histogram of the computed function may exhibit several maxima and minima, it is still meaningful to split it into two parts by means of a single threshold. Understandably, this is only an approximation and serves to present the potential of our approach. More sophisticated histogram analysis methods can be investigated in the future.

IV. Experiments

The proposed method has been tested on two datasets of different timbral characteristics. The first dataset (D_1) is the publicly available "Cante-100" collection of 100 flamenco recordings taken from commercially available flamenco anthologies [25]. The collection covers a large variety of singers and styles and encompasses a total of ≈ 6 hours of audio data. Vocal sections were manually annotated and cover approximately 55% of the total duration. The accompaniment instrumentation is limited to the guitar and rhythmic hand-clapping. Meta-data, audio descriptors and manual annotations are publicly available at www.cofla-project.com.

The second dataset (D_2) is based on a playlist of 19 Greek folk music tracks that are available on YouTube[1]. The total duration of D_2 is 62.74 minutes. The tracks were annotated by the authors with respect to the presence of singing voice and the annotations are available via the COFLA project website[2].

In this second dataset, the singing voice is mainly accompanied by combinations of violin, lute, sandur and percussion, of which the violin may take over the main melodic line in the absence of the singing voice. A common trait of both datasets is that the texture of accompaniment does not exhibit abrupt changes throughout the recording from a spectral point of view. However, there are strong dynamic fluctuations in the accompaniment and loud sections might even "cover" the singing voice temporarily.

For the sake of comparison, we have reproduced the supervised method in [6] which is based on trained Gaussian Mixture Models. In the sequel, we will refer to this method as *Song-2013-GMM*. We selected this method because it is based on the easily reproducible standard approach of Gaussian mixture modelling, while taking into account the fact that despite the wealth of published approaches on singing voice detection, publicly available code is still not common practice for this task. Method *Song-2013-GMM* was tested using a 10-fold validation scheme on each dataset. The resulting frame-wise Precision (Pr), Recall (R) and F-measure (F) for the class of the singing voice are shown in Table I. The most notable observation is that performance drops significantly on D_2 which can be attributed to the more complex instrumentation of this dataset.

TABLE I: Performance of the *Song-2013-GMM* method

	Precision (%)	Recall (%)	F-measure (%)
D_1	92.01	84.78	88.25
D_2	82.56	81.59	82.07

To evaluate our approach, we performed experiments with two different features, i.e., the bark band representation of the DFT spectrum and the Mel Frequency Cepstrum Coefficients (MFCCs). The latter have been very successful in singing voice detection compared to other features [2]. We also experimented with the augmentation of the feature space with the delta coefficients of these features. The feature extraction stage was executed using a short-term window technique, where the moving window length and step were set equal to 2048 and 512 samples, respectively, for a sampling rate of 44100Hz. For the first feature, we used 24 bark bands [26] and for the second feature, we adopted 13 MFCCs stemming from a typical filter bank of 40 triangular filters [26]. In all cases,

[1] www.youtube.com/watch?v=vi4s9mz3ajQ&list=PL725882B55F451A9C
[2] www.cofla-project.com/eusipco2016/eusipco2016_submission.zip

we used dictionaries of 32 and 64 atoms with sparsity ranging from 1 to 8 atoms, to preserve reasonable execution times. It has to be noted that in this study we are not aiming at reporting results on a large number of feature combinations. We are rather focusing on exhibiting the potential of our approach on accompanied singing, hence the choice of the above features. Of course, an exhaustive study of features can be the objective of future research.

Figures 3 and 4 present the adopted frame-wise performance measures for the case of the bark band feature on the two datasets. It can be observed that, in general, dictionaries of 64 atoms outperform dictionaries of 32 atoms. Furthermore, the best performance, both on D_1 and D_2, is achieved by a 2-sparse representation and a dictionary of 64 atoms. The corresponding values of Precision, Recall and F-measure are shown in Table II. These values demonstrate that the proposed method is competitive to a standard supervised scheme (Table I) without the need for a training stage on labelled data.

TABLE II: Best performance of the proposed method with respect to the F-measure: bark bands, 2-sparse representation, 64 atoms

	Precision (%)	Recall (%)	F-measure (%)
D_1	93.95	76.95	84.6
D_2	85.95	79.04	82.35

When MFCCs were used, a performance drop was made evident (see Figure 5). Furthermore, when the feature space was augmented with delta coefficients, the bark band feature retained good performance but the performance in the case of the MFCCs dropped even more, mainly due to poor recall, as it can be seen in Table III, which contains a summary of the best obtained results for all dataset-feature combinations.

Table III also reveals that a dictionary of 32 atoms can sometimes outperform a larger dictionary of 64 atoms but it has to be noted that the performance difference is marginal because the whole procedure depends heavily on the position of a single threshold. In other words, when a histogram presents several modalities (peaks), small threshold shifts can affect the recall of the method and more sophisticated thresholding techniques are needed.

TABLE III: Performance summary of the proposed method. Boldfaced entries indicate best achieved performance

	Feat.	P. (%)	R. (%)	F. (%)	Dict. size	Spars.
D_1	Bark	**93.95**	76.95	**84.6**	64	2
	Bark+delta	92.10	**77.70**	84.28	64	2
	MFCCs	85.98	71.01	77.78	32	2
	MFCCs+delta	88.21	65.60	75.24	32	1
D_2	Bark	**85.95**	79.04	**82.35**	64	2
	Bark+delta	82.26	80.15	81.19	32	2
	MFCCs	62.70	**84.60**	72.02	64	2
	MFCCs+delta	61.44	47.79	53.76	32	3

V. Conclusions

We presented an unsupervised signing voice detector that is based on Dictionary Learning and performs competitively to a standard supervised scheme when evaluated on accompanied singing in music recordings. The contribution of our approach lies in the fact that the data to be segmented are treated as training data and are used to learn a reasonably small dictionary. The frequency of excitation of the atoms of this dictionary, in conjunction with their intensity of contribution during signal reconstruction, can serve to compute the value of a weight function for each frame of the signal, the histogram of which can be used to compute a global threshold for the binarization of the feature sequence. Our future research will reveal if a more elaborate analysis of the modalities present in the histogram can yield better threshold estimates.

Acknowledgment

This work has been partly funded by the Junta de Andalucía, project COFLA II (P12-TIC-1362) and it has been partly supported by the University of Piraeus Research Center.

References

[1] T. L. Nwe and H. Li, "On fusion of timbre-motivated features for singing voice detection and singer identification," in *Proc. of IEEE ICASSP*, March 2008, pp. 2225–2228.

[2] M. Rocamora and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," in *Proc. of the 11th Brazilian Symposium on Computer Music*, São Paulo, Brazil, Sep. 2007.

[3] H. Lukashevich, M. Gruhne, and C. Dittmar, "Effective singing voice detection in popular music using arma filtering," in *Proc. of DAFx-07, Bordeaux, France*, 2007.

[4] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," in *Proc. of IEEE ICASSP*, April 2009, pp. 1685–1688.

[5] V. Rao, S. Ramakrishnan, and P. Rao, "Singing voice detection in polyphonic music using predominant pitch." in *Proc. of INTERSPEECH*, 2009, pp. 1131–1134.

[6] L. Song, M. Li, and Y. Yan, "Automatic vocal segments detection in popular music," in *Proc. of the Ninth International Conference on Computational Intelligence and Security*, 2013.

[7] V. Rao, C. Gupta, and P. Rao, "Context-aware features for singing voice detection in polyphonic music," in *Adaptive Multimedia Retrieval. Large-Scale Multimedia Retrieval and Evaluation*. Springer, 2011.

[8] B. Lehner, R. Sonnleitner, and G. Widmer, "Towards light-weight, real-time-capable singing voice detection," in *Proc. of ISMIR*, 2013.

[9] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. of ISMIR*, 2005.

[10] S. Santosh, S. Ramakrishnan, V. Rao, and P. Rao, "Improving singing voice detection in presence of pitched accompaniment," in *Proc. of the National Conference on Communications (NCC)*, 2009.

[11] H. Lachambre, R. André-Obrecht, and J. Pinquier, "Singing voice detection in monophonic and polyphonic contexts," in *Proc. of EUSIPCO*, 2009, pp. 1344–1348.

[12] N. Cho and C.-C. Kuo, "Sparse music representation with source-specific dictionaries and its application to signal separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 326–337, Feb 2011.

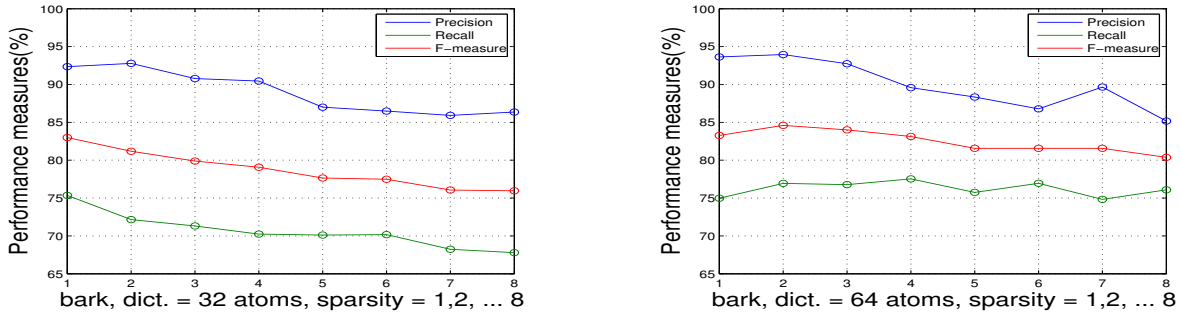
[13] B. L. Sturm and P. Noorzad, "On automatic music genre recognition by sparse representation classification using auditory temporal modulations," *Computer music modeling and retrieval*, pp. 379–394, 2012.

[14] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1905–1917, Dec. 2014.

[15] M. Srinivas, D. Roy, and C. Mohan, "Learning sparse dictionaries for music and speech classification," in *Proc. of the 19th International Conference on Digital Signal Processing (DSP)*, 2014.

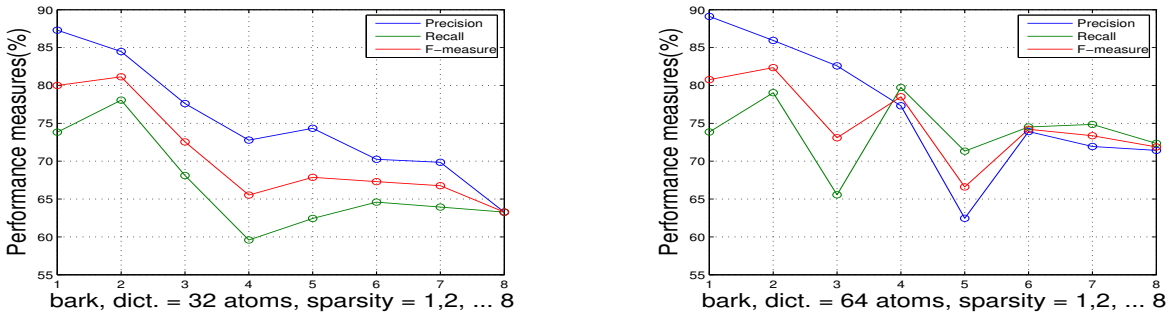
[16] J. Nam et al., "Learning sparse feature representations for music annotation and retrieval." in *Proc. of ISMIR*, 2012, pp. 565–570.

[17] C.-T. Lee, Y.-H. Yang, and H. H. Chen, "Multipitch estimation of piano music by exemplar-based sparse representation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 608–618, 2012.



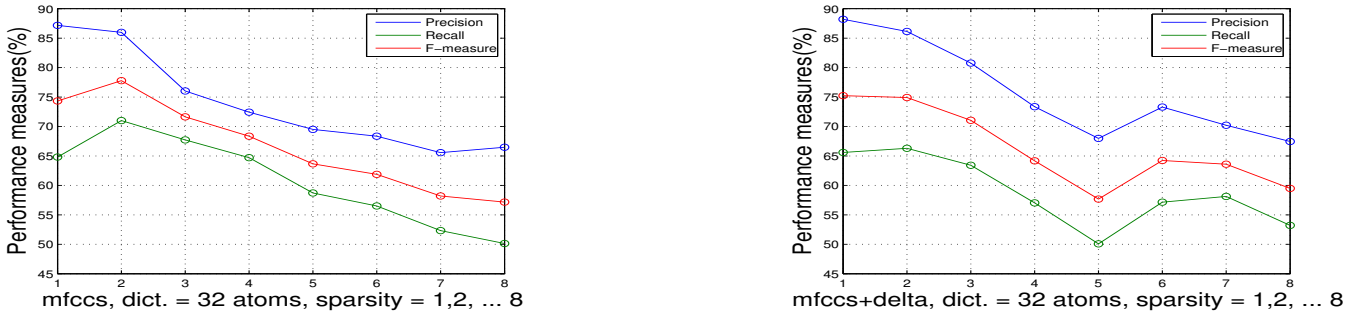
(a) Dictionary of 32 atoms (b) Dictionary of 64 atoms

Fig. 3: Precision Recall, and F-measure on D_1 using the bark-band feature



(a) Dictionary of 32 atoms (b) Dictionary of 64 atoms

Fig. 4: Precision Recall, and F-measure on D_2 using the bark-band feature



(a) Dictionary of 32 atoms (b) Dictionary of 32 atoms

Fig. 5: Precision Recall, and F-measure on D_1 using MFCCs and MFCCs augmented with delta coefficients.

[18] K. O'Hanlon, H. Nagano, and M. Plumbley, "Structured sparsity for automatic music transcription," in *Proc. of IEEE ICASSP*, March 2012, pp. 441–444.

[19] J. J. Carabias-Orti et al., "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1671–1680, 2013.

[20] T. B. Yakar et al., "Bilevel sparse models for polyphonic music transcription." in *Proc. of ISMIR*, 2013, pp. 65–70.

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[22] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[23] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[24] S. Theodoridis, Y. Kopsinis, and K. Slavakis, *Sparsity-aware learning and compressed sensing: An overview*. Academic press, 2014.

[25] N. Kroher, J. M. Díaz-Báñez, J. Mora, and E. Gómez, "Corpus cofla: A research corpus for the computational study of flamenco music," *ACM Journal on Computing and Cultural Heritage (in print)*, 2016.

[26] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Prentice-Hall Inc., 2011.