SINGING / RAP CLASSIFICATION OF ISOLATED VOCAL TRACKS

Daniel Gärtner
Fraunhofer Institute for Digital Media Technology IDMT
daniel.gaertner@idmt.fraunhofer.de

ABSTRACT

In this paper, a system for the classification of the vocal characteristics in HipHop / R&B music is presented. Isolated vocal track segments, taken from acapella versions of commercial recordings, are classified into classes singing and rap. A feature-set motivated by work from song / speech classification, speech emotion recognition, and from differences that humans perceive and utilize, is presented. An SVM is used as classifier, accuracies of about 90% are achieved. In addition, the features are analyzed according to their contribution, using the IRMFSP feature selection algorithm. In another experiment, it is shown that the features are robust against utterance-specific characteristics.

1. INTRODUCTION

According to the IFPI Digital Music Report 2010 [11] , the catalogue of digital music from the licensed music services contained more then 11 million tracks in 2009. For some years, researchers have been working on tools that simplify the handling of this large amount of data. Automatic content-based analysis is now part of a multitude of different applications. The algorithms help people to visualize their music collections and generate playlists. Music lovers can discover new music with the help of music recommendation engines. DJs use software for automatic tempo and beat detection.

This work is about automatically labeling short snippets of isolated vocal tracks according to their vocal characteristics. The segments are classified into two classes, rap and singing. These two classes are the dominant vocal styles in HipHop and contemporary R&B music. A successful labeling could further be useful in urban sub-genre classification, serve as a basis for vocal characteristics song segmentation, and help analyzing the song structure. Also, intelligent audio players could be designed, that automatically skip all sung or all rapped parts in R&B and HipHop music songs, depending on the preferences of their users.

Rap is a form of rhythmically speaking, typically to accompaniment music. As pointed out in [7], singing contains a larger percentage of voiced sounds than speaking. For Western music, the singing voice also covers a wider range of fundamental frequencies. In addition, singing tends to have a much wider dynamic range in terms of amplitude. According to [8], singing voice tends to be piecewise constant with abrupt changes of pitch in between. In natural speech, the pitch frequencies slowly drift down with smooth pitch change in an utterance. This peculiarity can also often be observed in rap passages. While rapping, artists are quite free in their choice of the pitches, while the fundamental frequencies in singing are usually related to the harmonic or melodic structure of the accompaniment.

In a survey conducted in [5], subjects had to label vocal utterances with a value from 1 (speaking) to 5 (singing), and explain their decision. For one utterance, 5 subjects used the word ”rap” in their explanation. The mean score of this utterance was 3.74. Rap seems to be perceived somewhere in between singing and speaking, in that special case even a bit more singing than speaking. Different subjects mentioned melody, rhythm, or rhyming combined with musical scales as features to discriminate singing from speaking. However, rhythm descriptions might be less important for rap / singing classification, since rap and singing are both rhythmically while speech is not. Further, repetitions, the clarity and sharpness of pitch, or the presence of vibrato have been identified to be present in singing rather than speaking. Another feature for the discrimination of speech and song as denoted in [9] is stress. It is stated that in English language speech, stress affects the meaning of the utterance. This is another one of the points where speech and rap differ. In rap, where the voice is used as instrument, accentuation often is part of the rhythm.

In previous work [4] the classification into singing and rap has been investigated on full songs (vocals + accompaniment), using common low-level features and a Gaussian mixture model based classifier. One of the outcomes of this work has been, that, although the classifier produced reasonable results, the classification was highly influenced by the accompaniment music. We therefore suggest to build the system composed of two major components: vocal track isolation and the classification of isolated tracks, using a feature set designed towards this task. This paper focuses on the second objective.

To the knowledge of the authors, automatic content-based discrimination of isolated singing and rap tracks has not yet been investigated elsewhere. However, research has been carried out on the task of singing and speaking classification. Investigations on the rap voice in a musicology context have been carried out though, e.g., [6].

In [5], a set of features is presented to discriminate between singing and speaking including statistics over F0 and ΔF0, vibrato detection, repetition detection, and the proportions of voiced frames, unvoiced frames, and silent frames, and repetition features.

Another system is presented in [19]. Based on features like the rate of voiced sounds, the standard deviation of the duration of voiced sounds, and the mean silence durations, an SVM is trained for singing / speaking classification. This classifier is used to forward sung queries to a query by humming system, and spoken queries to a speech recognition system.

[12] present another study on speaking / singing discrimination. The first part addresses human performance. They find that already 1 second of audio signal is enough to classify with an accuracy of 99.7% accuracy. Still 70% are reported on signals of 200 ms length. Further, it is investigated, that the performance drops, when either spectral or prosodic information is distorted in the audio signal. In the second part, the task is performed using Mel frequency cepstral coefficients (MFCCs), ΔMFCCs, and ΔF0 as features and a maximum likelihood classifier based on Gaussian mixture models (GMM).

Another field working with energy-based and pitch-features on vocal signals is speech emotion recognition (e.g., [17, 18]).

The remainder of this paper is organized as follows. In Section 2, the features and classifier are described. Section 3 deals with the experiments that have been conducted. There, also the used data and the general experimental setup is introduced. The results and their meaning are discussed in Section 4. Finally, the conclusions and an outlook are given in Section 5.

2. APPROACH

In this section, the used features and the classifier that has been utilized, are explained.

2.1 Features

The features contain the information about the audio signal, that is accessed by the classifier. Therefore, it is important, that the features are well designed with respect to the task.

Some of the features are calculated from the pitch of the vocal segment. YIN [3] has been used as F0-estimator. In addition to an F0-estimation in octaves over time, YIN's output also includes the instantaneous power (IP) and the ratio of aperiodic power to the total power (ATR).

All F0-estimations are transformed in the relative pitch representation (RPR), which is a mapping into an interval of one octave width around the dominant frequency. First, a histogram with 100 bins is calculated over the estimated F0 values. The center frequency of the bin with the highest value in the histogram is used as dominant frequency. Too large or small frequencies are halved or doubled respectively, until they fit into the chosen interval. By doing so, octave-errors are removed. Of course, also absolute pitch information is removed, but absolute pitch is mainly artist depended, and a contribution to rap / singing classification is not expected. The resolution of the YIN features is 1378 samples per second. Figure 1 and Figure 2 show the

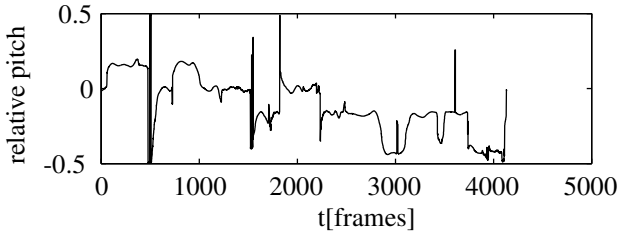

Figure 1. RPR progression of a singing snippet.

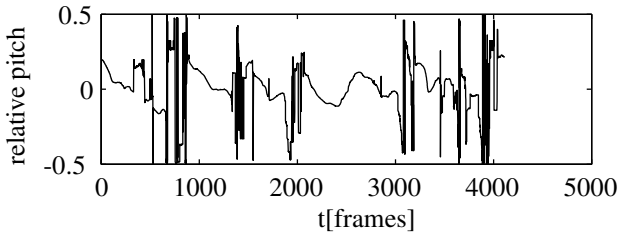

Figure 2. RPR progression of a rap snippet.

RPR progression for an exemplary singing and rap snippet respectively. One difference between the two examples is, that for singing, regions of almost constant pitch (RPR values of approximately 0.2, 0.0, -0.2, and -0.45 in Figure 1) can be observed, while for rap the RPR values are permanently changing. Based on RPR, IP, and ATR, a set of features is extracted.

First of all, the number of non-silent frames is determined, based on thresholding of IP. The ratio of non-silent frames to the number of overall frames will be denoted **ERatio**.

Next, from the non-silent frames, the number of voiced frames is determined, using a threshold on ATR. The ratio of voiced frames to the number of non-silent frames will be denoted **VRatio**. As already stated, rap is supposed to contain less voiced frames than song.

In another in-house study it has been discovered, that song segments have a lower syllable-density than rap segments. IP can be used as onset detection function. Based on adaptive thresholding, the number of onsets is estimated, which is then divided by the length of the segment. This feature is denoted **ORatio**.

As another step, from the voiced frames the segments are determined, during which $|\Delta\text{RPR}|$ is below a threshold. Segments of a length smaller then 10 frames are discarded. The ratio of frames that contribute to such a segment and the number of voiced frames is denoted **CRatio**. All the following calculations are performed on the RPR frames, that belong to a segment.

The mean of ΔRPR and the mean of ΔΔRPR also serve as features, denoted **PitchDiff** and **PitchDDiff**. Further, the mean of $|\text{RPR}|$, **MeanC**, and the variance of RPR, **VarC** are calculated.

The ratio of the number of frames with negative ΔRPR and the number of frames with positive ΔRPR is denoted **SLRatio**. In sung segments, either constant or with vibrato, both components are balanced. However, in rap segments a decreasing pitch can often be observed, and as a consequence, the SLRatio would be larger than 1.

A histogram over RPR with a resolution of 3 bins for each note is calculated, for a coarse approximation of the shape of the pitch distribution. Rap segments tend to have an unimodal RPR-distribution (Figure 3). Sung segments often have multimodal RPR-distributions, depending on the number of different notes that are sung in an utterance, as depicted in the example of Figure 4. Further, the RPR-distribution of a sung segment tends to have much sharper peaks than the distribution of a rap segment. The distance of the two bins with the largest values, divided by the width of the histogram will be denoted **NoteDist**. Dividing the second largest value in the histogram by the largest one, leads to the **NRatio**.

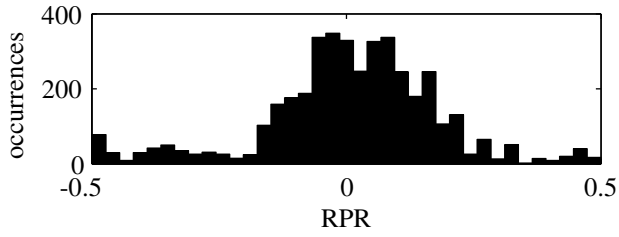

Figure 3. RPR-histogram for a rap snippet.

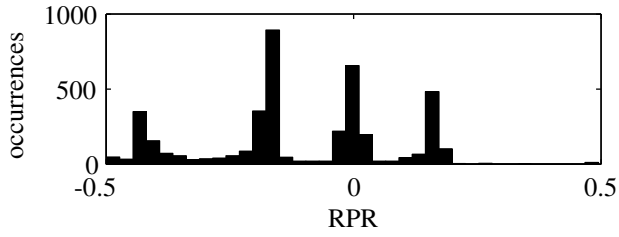

Figure 4. RPR-histogram for a singing snippet.

In addition, MFCCs are extracted from the audio signal. MFCCs are a popular feature in speech recognition and describe the spectral envelope. In [10], their applicability to modeling music has been shown, and as a consequence they have been successfully used in different music information retrieval tasks since then. For each snippet, the mean of all contributing frames is calculated. MFCCs are not part of the feature-set, they are used in a system for comparison reasons to describe the robustness of the feature-set in terms of utterance-sensitivity.

2.2 Classifier

Support vector machines (SVM, [2]) have been used as classifier. An SVM consists of a set of support vectors, that span a hyperplane in the feature space. This hyperplane separates two classes. The class of a test-observation depends on on which side of the hyperplane the test-observation is located in the feature-space. This can be calculated incorporating the dot-products of the feature vector and the support vectors. In the training stage, the support vectors are determined based on training observations.

In order to use non-linear hyperplanes, the feature space is transformed in a higher-dimensional space by the use of a kernel function. Computational costs for the transformation and the calculation of the dot-products can be reduced by selecting the kernel in a way that the dot-product can also be expressed in the original feature space. A radial basis function (RBF) kernel has been used, that is parameterized by γ.

Another parameter of the SVM is C, the weight of the error term during training. LibSVM [1] has been used as SVM implementation.

3. EXPERIMENTS

Following the approach section, the system setup, including the data, and the performed experiments are explained.

3.1 Data

A dataset of 62 songs from 60 different artists has been used in this study. Acapella versions of commercial Hip-Hop and contemporary R&B songs, performed in English language have been used. In these genres, songs are often released including an acapella and instrumental version. Other artists or DJs then can make remixes. For all songs, the segments containing only monophonic singing or monophonic rap have been determined.

Each segment is cut into 3s snippets, that overlap by 0.5s. The influence of the segment length is not evaluated in this paper. Although [12] reports that already snippets of 1 second length contain enough information for humans to accurately classify speech and singing, a larger snippet size has been chosen, since it is then more likely to observe notes with different pitches in singing snippets. The final dataset consists of 815 rap-snippets and 584 singing-snippets.

3.2 System setup and evaluation

Training and evaluation is performed using 5-fold cross-validation. All snippets are randomly distributed amongst the 5 folds using an utterance filter, which means, that all snippets from one song (belonging to one utterance) are distributed in the same fold. Each of the folds serves as test-data once and is part of the training-data in the other cases. The training data is used to determine the parameters of the SVM, i.e., the support vectors, C, and γ. It is crucial in SVM training / classification that all the features have approximately the same range. Therefore the data has to be normalized. Variance-normalization is used, in order to make the data zero mean unit variance. The mean μ and the standard deviation σ have to be estimated.

A reasonable choice of C and γ is important for good classification results. Both parameters are estimated using 3-fold cross-validation on the training data. This stage

will later be referred to as development stage. The distribution into the folds is done randomly again. However, at this point it is possible to decide whether an utterance filter should be applied or not. A three-stage grid-search has been employed. Since this process itself also consists of training and evaluation, μ and σ have to be determined every time the training-data changes due to recomposition from different folds.

Having determined C and γ, μ and σ are estimated based on the whole training-data, the training-data is normalized, and the SVM is trained with the previously determined C and γ. Finally, the performance is measured using the test-data, which is the test-observations from the one out of five folds, that has not been used for training and development.

The performance of a trained system both in evaluation A_t and development A_d is measured in accuracy. The accuracy of a classifier on given data is calculated by dividing the number of correctly classified test-observations by the number of all test-observations. Accuracy can be sensitive to imbalanced test-data. So if for example the test-data contains 80% observations from one class and only 20% observations from the other class, a classifier, that would always choose the same class would lead to a performance of either 80% or 20%, depending on which class he always choses. Therefore the test data is made balanced during the evaluation by randomly picking 584 observations from the 815 rap snippets.

The whole process, incorporating the random distribution into five folds, the development and training of the classifier, and its evaluation, is performed multiple times (denoted #runs), since this process contains random elements and is therefore indeterministic. The mean and variance of the accuracies in the test series are given as final measure, denoted $\mu_{A,t}$ and $\sigma^2_{A,t}$. Further, in Table 2 also $\mu_{A,d}$ and $\sigma^2_{A,d}$, are given, which are the accuracies during development for the chosen C and γ. Matlab is used as experimental framework.

3.3 Feature selection

A feature selection algorithm (FSA) has been used to give an estimation of the contribution of each of the features. Inertia ratio maximization using feature space projection [16] is a filter FSA, where the criteria of choosing features is distinct from the actual classifier. For each feature dimension an r-value is determined, which is the ratio of the between-class inertia to the total-class inertia. The feature with the largest r is chosen, then an orthogonalization process is applied to the feature space, in order to avoid the choice of redundant dimensions during following iterations. These steps are repeated until a stop criterium applies. The order of the features after feature selection reflects their importance according to the feature selection criterion, that should be correlated to the classification performance to a certain extend.

3.4 Utterance filter

One of the goals in machine learning is to build systems that are able to generalize. Also, performances of classifiers should be compared based on unseen test data. In order to achieve this, it is necessary to strictly discriminate training-data and testing-data during development and evaluation of the system. The distribution of the data in training and test-set can be even more restricted. It is common practice to put all the segments of a song in the same dataset, to for example avoid that the system is trained with a segment from the song and tested with a similar segment from the same song. In [15], it is suggested to put all pieces of an artist in the same dataset in a genre classification task. With experiments it is shown, that the performance of a system decreases significantly, if this so called artist filter is used. A possible reason for this is, that the system might focus on perceptually not so relevant information such as production effects [14].

As described in 3.2, an utterance filter is always applied in the 5-fold cross-validation setup, since it is possible, that the suggested feature set also reflects utterance-specific characteristics. In the 3-fold cross-validation development stage however, the utterance-filter can be either applied or omitted. Comparing performances based on systems with and without utterance-filter helps in describing the robustness towards utterance-specific characteristics. If a system generalizes well, $\mu_{A,t}$ and $\mu_{A,d}$ should be approximately equal.

The mean over the MFCC-frames of a snippet is a feature, that is supposed to be utterance-specific. In 4.3, the use of an utterance-filter is analyzed for the proposed feature-set and the mean-MFCC feature.

4. RESULTS AND DISCUSSION

The results of the performed experiments are listed and discussed in this section.

4.1 Feature contribution

In Table 1, the outcome of the FSA is denoted. Overall, feature selection has been performed 69425 times. In all runs, the VRatio feature has been selected first, as can be seen in column 2, belonging to rank 1. Further important features are CRatio, SLRatio and ORatio, that have been chosen 54989, 8175, and 6240 times as second feature respectively. The most unimportant features according to the IRMFSP are PitchDDiff and VarC (often chosen on rank 10 or 11 according to the values in column 11 and column 12).

The mean r-value of the first selected feature is 0.52, followed by 0.47 for the second selected feature. r decreases drastically from the second to the third selected feature. In [16], it is suggested to stop the iterative feature selection process as soon as r of the current iteration is below 1/100 of r in the first iteration. Following this criterion, the 6 top features would have been selected.

Rank	1	2	3	4	5	6	7	8	9	10	11
$\bar{r}$	0.5191	0.4716	0.0459	0.0151	0.0094	0.0068	0.0046	0.0034	0.0018	0.0004	0.0001
CRatio	0	54989	0	2	91	54	891	4511	5932	1678	1277
ERatio	0	0	0	359	2218	6062	13300	24101	17819	3644	1922
MeanC	0	0	0	1367	21200	9070	8650	13060	12116	3257	705
NoteDist	0	21	6432	50987	8455	2720	686	120	4	0	0
NRatio	0	0	0	3505	12729	8077	12597	9916	21221	1305	75
ORatio	0	6240	1880	10822	11029	9791	20787	7385	1376	112	3
PitchDDiff	0	0	0	0	5	54	794	3593	7596	39962	17421
PitchDiff	0	0	12249	2342	12096	27695	9244	4585	1195	17	2
SLRatio	0	8175	48864	41	1601	5902	2464	1999	357	22	0
VarC	0	0	0	0	1	0	12	155	1809	19428	48020
VRatio	69425	0	0	0	0	0	0	0	0	0	0

Table 1. Ranks of different features in the feature selection process.

4.2 System Performance

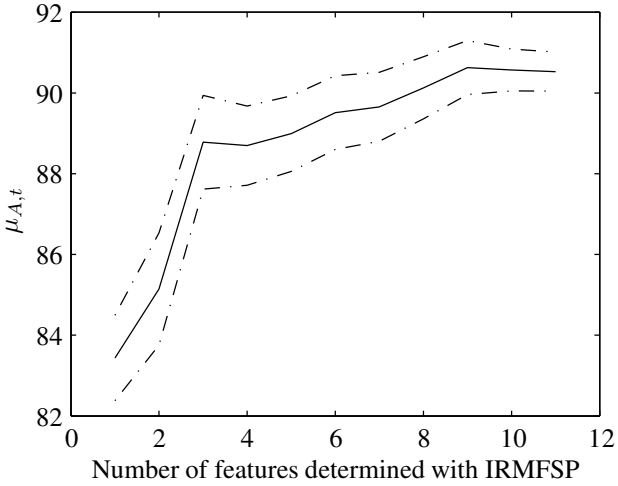

Figure 5. Performance subject to the number of features.

The final performance of the system is plotted against the number of features after IRMFSP in Figure 5. The top performance, 90.62% is achieved using 9 features. Using the feature-set consisting of all 11 features leads to a mean accuracy of 90.53%. The largest gain in performance is reported from 2 features (85.14%) to 3 features (88.78%).

4.3 Influence of the utterance filter

Table 2 contains the results of the investigation of utterance-sensitivity. For the suggested feature-set (full) the performance decrease is 1.09% (91.63% down to 90.54% from development to testing) with utterance filter. Without utterance-filter 3.07% (from 93.88% down to 90.81%) are observed. These small decreases originate in the fact that $\mu_{A,d}$ is result of the optimization of C and γ, while $\mu_{A,t}$ is not. Further, during development, imbalanced test-data is used for the evaluation, which can also lead to differences between both values. On the full feature-set, $\mu_{A,t}$ is almost similar for both systems, with and without utterance filter.

Feat.	u.filter	$\mu_{A,t}$	$\sigma^2_{A,t}$	$\mu_{A,d}$	$\sigma^2_{A,d}$	#runs
full	yes	90.54	0.53	91.63	0.16	1588
full	no	90.81	0.49	93.88	0.03	1583
MFCC	yes	67.71	6.71	72.84	1.17	1084
MFCC	no	65.08	3.85	96.36	0.04	934

Table 2. Influence of the utterance-filter.

Applying an utterance-filter to the MFCC-feature results to an decrease from 5.13% (from 72.84% down to 67.71%), which again can be explained with the optimization procedure. If the system is trained with the MFCC-feature without using an utterance-filter, the development-performance is 96.36%, which is the highest one achieved in the experiments. But on new utterances, the performance drastically decreases to 65.08%. In our data, artists that rap do not sing and vice versa. Without the utterance-filter, different parts of the same utterance are in the test-set and the training-set during the system-development, and a task like that can also be performed by an artist-detection or utterance-detection system. MFCCs are well known for their capabilities to capture speaker characteristics, and are therefore often used in speaker recognition systems. So in the development stage, the system is trained to classify into rap and singing by actually identifying utterances. A $\mu_{A,d}$-value of 96.36% shows, that, MFCCs are an appropriate feature for this task. On the contrary, $\mu_{A,t}$ is determined classifying snippets from unknown utterances. An utterance detection system cannot do that well, which leads to a low accuracy of 65.08%. For the MFCC-system with utterance filter, as already reported the difference is much smaller. For the full feature-set, no large difference between $\mu_{A,t}$ and $\mu_{A,t}$ could be observed. This set therefore is not sensitive to utterance-specific characteristics.

Comparing $\mu_{A,t}$ for the MFCC-systems with and without utterance-filter, one can see that the system trained with utterance-filter performs 2.63% better. A possible reason is that MFCCs seem to be able to also classify based the vocal characteristics to a certain extend, but when trained

without utterance-filter, the classifier seems to "learn the task that is easier to perform", which might be utterance-identification instead of vocal characteristics classification. When trained with utterance-filter, there is no utterance-identification development data provided. But since the difference is so small, there might be other reasons.

5. CONCLUSIONS AND OUTLOOK

A system for the classification of isolated vocal tracks into the classes singing and rap has been presented. A feature set, motivated by differences perceived by human is developed. Accuracies of over 91% are achieved on 3 second snippets of isolated vocal tracks from commercial urban music recordings. Further, it has been shown in experiments with an utterance-filter, that the suggested feature-set is not sensitive to utterance-specific characteristics.

As a next step, the application on full tracks, where no isolated vocal tracks are available, will be investigated. Since the described system is not designed to also work on mixtures of vocal tracks and accompaniment, the vocal track has to be separated from the song. Methods for the separation of the vocal track as described in, e.g., [13, 20, 21] are currently investigated. The system that has been described in this paper can also serve as benchmark for the source separation algorithms. Further, a study incorporating listening test is intended, in order to evaluate human performance on this task.

6. REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273:297, 1995.

[3] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America*, 111 (4):1917:1930, 2002.

[4] Daniel Gärtner and Christian Dittmar. Vocal characteristics classification of audio segments: An investigation of the influence of accompaniment music on low-level features. In *Proceedings of the ICMLA*, 2009.

[5] David Gerhard. *Computationally measurable differences between speech and song*. PhD thesis, Simon Fraser University, Canada, 2003.

[6] Ferdinand Hörner and Oliver Kautny. *Die Stimme im HipHop*. transcript Verlag, 2009.

[7] Youngmoo E. Kim. *Singing Voice Analysis/Synthesis*. PhD thesis, Massachusetts Institute of Technology, 2003.

[8] Yipeng Li and DeLiang Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1475–1487, May 2007.

[9] George List. The boundaries of speech and song. *Ethnomusicology*, 7(1):1:16, January 1963.

[10] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of ISMIR*, 2000.

[11] International Federation of the Phonographic Industry. IFPI Digital Music Report 2010. Available at http://www.ifpi.org/content/library/DMR2010.pdf.

[12] Yasunori Ohishi, Masataka Goto, Katunobu Itou, and Kazuya Takeda. On human capability and acoustic cues for discriminating singing and speaking voices. In *Proceedings of ICMPC*, 2006.

[13] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval. Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, July 2007.

[14] Elias Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval*. PhD thesis, Vienna University of Technology, Austria, March 2006.

[15] Elias Pampalk, Arthur Flexer, and Gerald Widmer. Improvements of audio-based music similarity and genre classification. In *Proceedings of ISMIR*, London, UK, 2005.

[16] Geoffroy Peeters and Xavier Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. In *Proceedings of DAFx*, 2003.

[17] Thomas S. Polzin. Verbal and non-verbal cues in the communication of emotions. In *Proceedings of ICASSP*, 2000.

[18] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *Proceedings of ICASSP*, 2003.

[19] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Discrimination of speech and monophonic singing in continuous audio streams applying multi-layer support vector machines. In *Proceedings of ICME*, volume 3, pages 1655–1658, 2004.

[20] Shankar Vembu and Stephan Baumann. Separation of vocals from polyphonic audio recordings. In *Proceedings of ISMIR*, 2005.

[21] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *Proceedings of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2008.