TIMBRE AND MELODY FEATURES FOR THE RECOGNITION OF VOCAL ACTIVITY AND INSTRUMENTAL SOLOS IN POLYPHONIC MUSIC

Matthias Mauch Hiromasa Fujihara Kazuyoshi Yoshii Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{m.mauch, h.fujihara, k.yoshii, m.goto}@aist.go.jp

ABSTRACT

We propose the task of detecting instrumental solos in polyphonic music recordings, and the usage of a set of four audio features for vocal and instrumental activity detection. Three of the features are based on the prior extraction of the predominant melody line, and have not been used in the context of vocal/instrumental activity detection. Using a support vector machine hidden Markov model we conduct 14 experiments to validate several combinations of our proposed features. Our results clearly demonstrate the benefit of combining the features: the best performance was always achieved by combining all four features. The top accuracy for vocal activity detection is 87.2%. The more difficult task of detecting instrumental solos equally benefits from the combination of all features and achieves an accuracy of 89.8% and a satisfactory precision of 61.1%. With this paper we also release to the public the 102 annotations we used for training and testing. The annotations offer not only vocal/non-vocal labels, but also distinguish between female and male singers, and different solo instruments.

Keywords: vocal activity detection, pitch fluctuation, F0 segregation, instrumental solo detection, ground truth, SVM

1. INTRODUCTION

The presence and quality of vocals and other melody instruments in a musical recording are understood by most listeners, and often these are also the parts of the music listeners are interested in. Music enthusiasts, radio disk-jockeys and other music professionals can use the locations of vocal and instrumental activity to efficiently navigate to the song position they're interested in, e.g. the first vocal activity, or the guitar solo. In large music collections, the locations of vocal and instrumental activity can be used to offer meaningful audio thumbnails (song previews) and better browsing and search functionality.

Due to its apparent relevance to music listeners and in commercial applications the automatic detection of vocals in particular has received considerable attention in the recent Music Information Retrieval literature, which we review below. Far less attention has been dedicated to the detection of instrumental solos in polyphonic music recordings.

In the present publication we present a state-of-the-art method for vocal activity detection. We show that the use of several different timbre-related features extracted based on a preliminary extraction of the predominant melody line progressively improve the performance of locating singing segments. We also introduce the new task of instrumental solo detection and show that, here too, the combination of our proposed features leads to substantial performance increases.

Several previous approaches to singing detection in polyphonic music have relied on multiple features. Berenzweig [2] uses several low-level audio features capturing the spectral shape, and learned model likelihoods of these. Fujihara uses both [3] a spectral feature and a feature that captures pitch fluctuation based on a prior estimation of the predominant melody. Thus more aspects of the complex human voice can be captured and modelled. In fact, Regnier and Peeters [14] note that "the singing voice is characterized by harmonicity, formants, vibrato and tremolo". However, most papers are restricted to a small number of (usually spectral) features [8, 9, 14]. Nwe and Li [12] have proposed the most diverse set of features for vocal recognition that we are aware of, including spectral timbre, vibrato and a measure of pitch height.

Our method is similar to that of Nwe and Li in that we use a wide range of audio features. However, our novel measurement of pitch fluctuation (similar to vibrato) is tuning-independent and based on a prior extraction of the predominant melody. Furthermore, we propose two new features that are also based on the preliminary melody extraction step: the timbre (via Mel-frequency cepstral coefficients) of the isolated predominant melody, and the relative amplitude of the harmonics of the predominant melody.

The remainder of the paper is organised as follows: in Section 2 we describe the features used in our study. Section 3 describes a new set of highly detailed ground truth annotations for more than 100 songs published with this paper. The experimental setup and the machine learning tools involved in training and testing our methods are explained Section 4. The results are discussed in Section 5. Limitations of the present method and future directions are discussed in Section 6.

2. AUDIO FEATURES

This section introduces the four audio features considered in this paper: the standard MFCCs, and three features based on the extracted melody line: pitch fluctuation, MFCCs of the re-synthesized predominant voice, and the relative harmonic amplitudes of the predominant voice.

We first extract all features from each track at a rate of 100 frames per second from audio sampled at 16 kHz, then low-pass filter and downsample them to obtain features at 10 frames per second, which we use as the input to the training and testing procedures (Section 4).

2.1 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients [11] are a vector-shaped feature which has the desirable property of describing the spectral timbre of a piece of audio while being largely robust to changes in pitch. This property has made them the *de facto* standard input feature for most speech recognition systems. The calculation of MFCCs consists of a discrete Fourier transform of the audio samples to the frequency domain, applying an equally-spaced filter bank in the mel frequency scale (approximately linear in log frequency), and finally applying the discrete cosine transform to the logarithm of the filter bank output. Details are extensively covered elsewhere, see e.g. [13]. In our implementation, the hop size is 160 samples (10 ms), the frame size is 400 samples (a 512-point FFT was used with zero-padding) and the audio window used is a Hamming window.

2.2 Pitch Fluctuation

The calculation of *pitch fluctuation* involves three steps:

fundamental F0: estimate the fundamental frequency (F0) of the predominant voice at every 10ms frame using PreFEst [4], and take the logarithm to map them to pitch space,

tuning shift: infer a song-wide tuning from these estimates, shift the estimates so that they conform to a standard tuning and wrap them to a semitone interval,

intra-semitone fluctuation: calculate the standard deviation of the frame-wise frequency difference.

We use the program PreFEst [4] to obtain an estimate of the fundamental frequency (F0) of the predominant voice at every 10ms frame. For a frame at position $t \in \{1, \ldots, N\}$ in which PreFEst detects any fundamental frequency $f[t]$ we consider its pitch representation $f^*_{\log}[t] = \log_2 f[t]$, i.e. the difference between two adjacent semitones is $\frac{1}{12}$.

The tuning shift in the second step is motivated as follows: our final pitch fluctuation measure employs pitch estimates wrapped into the range of one semitone. The wrapped representation has the benefit of discarding sudden octave jumps and similar transcription artifacts, but if the semitone boundary is very close to the tuning pitch of the piece, then even small fluctuations will cross this boundary (they 'wrap around') and lead to many artificial jumps of one semitone. This can be avoided if we shift the frequency estimates such that the new tuning pitch is at the centre of the wrapped semitone interval. In order to calculate the tuning of the piece we use a histogram approach (like [6]): all estimated values $f^*_{\log}[t], t \in \{1, \ldots, N\}$ are wrapped into the range of one semitone,

$$f^*_{\log}[t] \left(\mathrm{mod}\, \frac{1}{12} \right), t \in \{1, \ldots, N\}, \tag{1}$$

and sorted into a histogram $(h_1, \ldots, h_{100})$ with 100 histogram bins, equally-spaced at $\frac{1}{1200}$, or one cent. The relative tuning frequency is obtained from the histogram as

$$\begin{aligned} f^{\mathrm{ref}}_{\log} &= \frac{(\arg\max_i h_i) - 1}{1200} - 0.5 \\ &\in \{-0.5, -0.49, \ldots, 0.49\}, \end{aligned} \tag{2}$$

and the semitone-wrapped frequency estimates we use in the third step are

$$f_{\log}[t] = \left(f^*_{\log}[t] - f^{\mathrm{ref}}_{\log} \right) \left(\mathrm{mod}\, \frac{1}{12} \right), t \in \{1, \ldots, N\}.$$

The third step calculates a measure of fluctuation on windows of the frame-wise values $f_{\log}[t]$. We use Fujihara's formulation [3] of the frequency difference (up to a constant)

$$\Delta f_{\log}[t] = \sum_{k=-2}^{2} k \cdot f_{\log}[t+k] \tag{3}$$

and define pitch fluctuation as the Hamming-weighted standard deviation of values $\Delta f_{\log}[.]$ in a neighbourhood of t,

$$\mathrm{F}[t] = 12 \cdot \sqrt{\sum_{k=1}^{50} w_k \left(\Delta f_{\log}[t+k-25] - \mu[t] \right)^2}, \tag{4}$$

where $\mu[t] = \sum_{k=1}^{50} w_k \Delta f_{\log}[t+k-25]$ is the Hamming-weighted mean, and $w_k, k = 1, \ldots, 50$ is a Hamming window scaled such that $\sum_k w_k = 1$.

In short, $\mathrm{F}[t]$ summarises the spread of frequency changes of the predominant fundamental frequency in a window around the t^{th} frame.

2.3 MFCCs of Re-Synthesised Predominant Voice

We hypothesize that audio features that describe the predominant voice in a polyphonic recording in isolation will improve the characterisation of the singing voice and solo instruments. To obtain such a feature we re-synthesize the estimated predominant voice and perform the MFCC feature extraction on the resulting monophonic waveform. For the re-synthesis itself we use an existing method [3] which employs sinusoidal modelling based on the PreFEst estimates of predominant fundamental frequency and the estimated amplitudes of the harmonic partials pertaining to that frequency. MFCC features of the re-synthesized audio are calculated as explained in Section 2.1. They describe the spectral timbre of isolated the most dominant note.

2.4 Normalised Amplitudes of Harmonic Partials

The MFCC features described in Sections 2.1 and 2.3 capture the spectral timbre of a sound, but they do not contain information on another dimension of timbre: the normalised amplitudes of the harmonic partials of the predominant voice. Unlike the MFCC feature of the re-synthesised predominant voice, this feature uses the amplitude values themselves, i.e. at every frame the feature is derived from the estimated harmonic amplitudes $A = (A_1, \ldots, A_{12})$ by normalising them according to the Euclidean norm,

$$H_i = \frac{A_i}{\sqrt{\sum_i A_i^2}} \tag{5}$$

3. REFERENCE ANNOTATIONS

We introduce a new set of manually generated reference annotations to 112 full-length pop songs: 100 songs from the popular music collection of the RWC Music Database [5], and 12 further pop songs. The annotations describe activity in contiguous segments of audio using seven main classes: `f` – female lead vocal, `m` – male lead vocal, `g` – group singing (choir), `s` – expressive instrumental solo, `p` – exclusively percussive sounds, `b` – background music that fits none of the above, `n` – no sound (silence or near silence). There's also an additional `e` label denoting the end of the piece. In practice, music does not always conform to these labels, especially when several expressive sources are active. In such situations we chose to annotate the predominant voice (with precedence for vocals) and added information about the conflict, separated by a colon, e.g.

```
m:withf.
```

Similarly, the label for expressive instrumental solo, `s`, is always further specified by the instrument used, e.g.

```
s:electricguitar.
```

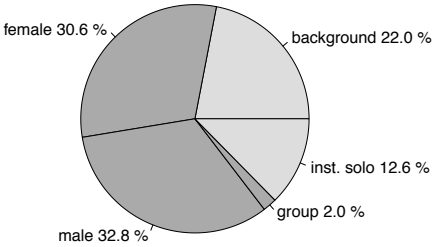


Figure 1: Ground truth label distribution: the pie chart labels provide information on the distribution in the *extended* model with five classes. The *simple* model joins all vocal classes (dark grey, 65.4%) and all non-vocal classes (light grey, 34.6%).

The reference annotations are freely available for download [1] .

4. EXPERIMENTS

We used 102 of the ground truth songs and mapped the rich ground truth annotation data down to fewer classes according to two different schemes:

simple contains two classes: *vocal* (comprising ground truth labels `f`,`m` and `g`) and *non-vocal* (comprising all other ground truth labels)

extended contains five classes: *female*, *male*, *group* for the annotations `f`,`m` and `g`, respectively; *solo* (ground truth label `s`); and *remainder* (all remaining labels)

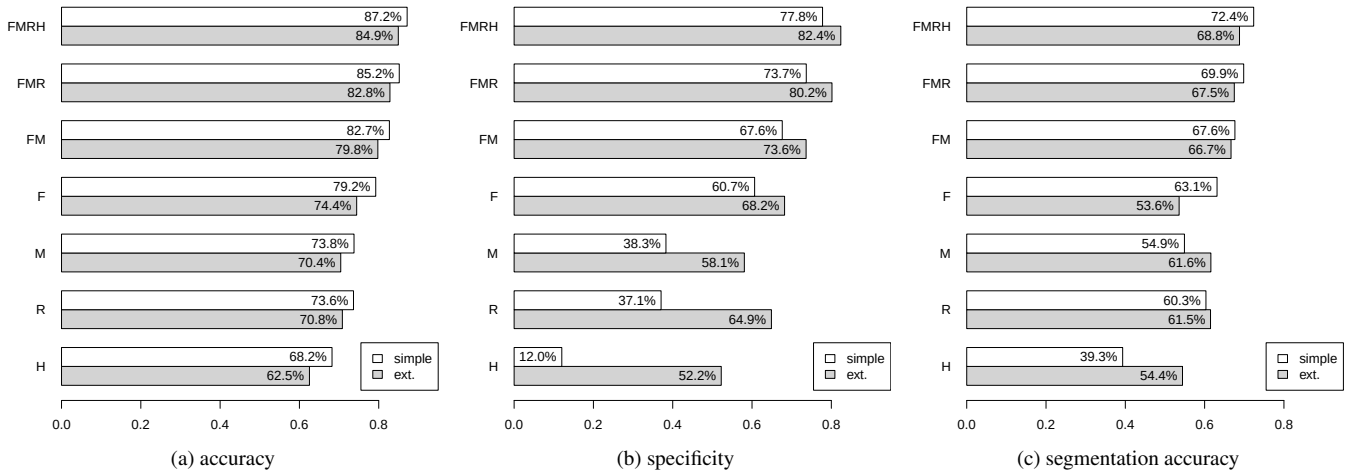
The frequency of the different classes is visualised in Figure 1. Short background segments (ground truth label `b`) of less than 0.5 s duration were merged with the preceding region.

We examine seven different feature configurations, the four single features pitch fluctuation (F), MFCCs (M), MFCCs of the re-synthesised melody line (R) and normalised aplitudes of the harmonics (H), and the following progressive combinations of the four: FM, FMR and FMRH.

The relevant features in each feature configuration are cast into a single vector per frame. We use the support vector machine version of a hidden Markov model [1] *SVM-HMM* [7] via an open source implementation [2] . We trained a model with the default order of 1, i.e. with the probability of transition to a state depending only on the respective previous state. The slack parameter was set to $c = 50$, and the parameter for required accuracy was set to $e = 0.6$. The 102 songs are divided into five sets for cross-validation. The estimated sequence is of the same format as the mapped ground truth, i.e. either two classes (*simple* schema) or five classes (*extended* schema).

[1] `http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/`

[2] `http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html`



(a) accuracy (b) specificity (c) segmentation accuracy

Figure 2: Vocal activity detection (see Section 5.1).

5. RESULTS

In order to give a comprehensive view of the results we use four frame-wise evaluation metrics for binary classification: accuracy, precision, recall/sensitivity and specificity. These metrics can be represented in terms of the number of true positives (TP; method says its positive and ground truth agrees), true negatives (TN; method says it's negative and ground truth agrees), false positives (FP; method says it's positive, ground truth disagrees) and false negatives (FN; method says it's negative, ground truth disagrees).

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{\# all frames}}, \qquad \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

We also provide a measure of segmentation accuracy as one minus the minimum of the directional Hamming divergences, as proposed by Christopher Harte in the context of measuring chord transcription accuracy. For details see [10, p. 52].

5.1 Vocal Activity Detection

Table 1 provides all frame-wise results of vocal activity detection in terms of the four metrics shown above. The highest overall accuracy of 87.2% is achieved by the *simple* FMRH method. The difference to the second-best algorithm in terms of accuracy (*simple* FMRH) is statistically significant according to the Friedman test (p value: $< 10^{-7}$).

Accuracy of single features. Figure 2a shows the distinct accuracy differences between the individual single audio features. The H feature by itself has a very low accuracy of 68.2% (62.5% in the extended model). The accuracy obtained by either the MFCC-based features, M and R are already considerably higher—up to 73.8%—and the pitch fluctuation measure F is the measure with the highest accuracy of 79.2% (73.4% in the extended model) among models with a single feature. This suggests that pitch fluctuation is the most salient feature of the vocals in our data.

Progressively combining features. It is also very clear that the methods using more than one feature have an advantage: every additional feature increases the accuracy of vocal detection. In particular, the R feature—MFCCs of the re-synthesised melody line—significantly increases accuracy when added to the feature set that already contains the basic MFCC features M. This suggests that R and M have characteristics that complement each other. More surprising, perhaps, is the fact that the addition of the H feature, which is a bad vocal classifier on its own, leads to a significant improvement in accuracy.

Precision and Specificity. If we consider the accuracy values alone it seems to be clear that the *simple* model is better: it outperforms the *extended* model in every feature setting. This is, however, not the conclusive answer. Accuracy tells only part of the story, and other measures such as precision and specificity are helpful to examine different aspects of the methods' performance. The recall measure does not provide very useful information in this case, because—unlike in usual information retrieval tasks—the *vocal* class occupies more than half the database, see Figure 1. Hence, it is very easy to make a trivial high-recall classifier by randomly assigning a high proportion x of frames to the positive class. To illustrate this, we have added theoretical results for the trivial classifiers 'rand-x' to Table 1. A more difficult problem, then, is to make a model that retains high recall but also has high precision and specificity. Specificity is the recall of the negative class, i.e. the ratio of *non-vocal* frames that have been identified as such, and precision is the ratio of truly *vocal* frames in what the automatic method claims it is. The *extended* methods outperform each corresponding *simple* method in terms of precision and specificity. Figure 2b also shows that better results are achieved

	accuracy	precision	recall	specificity
rand-0.500	0.500	0.654	0.500	0.500
rand-0.654	0.547	0.654	0.654	0.346
rand-1.000	0.654	0.654	1.000	0.000
simple H	0.682	0.678	0.979	0.120
simple R	0.736	0.736	0.930	0.371
simple M	0.738	0.739	0.926	0.383
simple F	0.792	0.811	0.891	0.607
simple FM	0.827	0.841	0.907	0.676
simple FMR	0.852	0.868	0.913	0.737
simple FMRH	**0.872**	0.887	0.921	0.778
ext. H	0.625	0.729	0.680	0.522
ext. R	0.708	0.799	0.740	0.649
ext. M	0.704	0.775	0.770	0.581
ext. F	0.744	0.822	0.777	0.682
ext. FM	0.798	0.856	0.830	0.736
ext. FMR	0.828	0.889	0.842	0.802
ext. FMRH	0.849	**0.903**	0.863	**0.824**

Table 1: Recognition measures for vocal activity.

by adding our novel audio features.

Segmentation accuracy. As we would expect from the above results, the segmentation accuracy, too, improves with increasing model complexity. The top segmentation accuracy of the top score of 0.724 is is approaching that of state-of-the-art chord segmentation techniques (e.g. [10, p. 88], 0.782). For the four best feature combinations the *simple* methods slightly outperform the *extended* ones, by 2 to 4 percentage points.

The best *extended* method, *extended* FMRH, has the highest precision (90.3%) and specificity (82.4%) values of all tested algorithms, while retaining high accuracy and recall (84.9% and 86.3%, respectively). In most situations this would be the method of choice, though the respective *simple* method has a slight advantage in terms of segmentation accuracy.

5.2 Instrumental Solo Activity

More difficult than detecting vocals is detecting the instrumental solos in polyphonic pop songs because they occupy a smaller fraction of the total number of frames (12.6%, see Figure 1). Hence, this situation is more similar to a traditional retrieval task (the desired positive class is rare), and precision and recall are the relevant measures for this task. Table 1 shows all results, and—for comparison—the theortical performance of the three classifiers 'rand-x' that randomly assign a ratio of x frames to the $solo$ class.

The method that includes all our novel audio features, FMRH, achieves the highest accuracy of all methods. However, all methods show high accuracy and specificity; precision and recall show the great differences between the methods. Figure 3 illustrates the differences in precision of solo

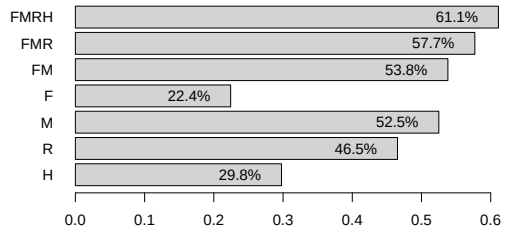


Figure 3: Detection of instrumental solos: precision of the *extended* methods.

	accuracy	precision	recall	specificity
rand-0.126	0.780	0.126	0.126	0.874
rand-0.500	0.500	0.126	0.500	0.500
rand-1.000	0.126	0.126	1.000	0.000
ext. H	0.829	0.298	0.262	0.911
ext. R	0.866	0.465	0.406	0.933
ext. M	0.877	0.525	0.290	0.962
ext. F	0.860	0.224	0.045	0.977
ext. FM	0.876	0.538	0.152	0.981
ext. FMR	0.889	0.577	0.445	0.953
ext. FMRH	0.898	0.611	0.519	0.952

Table 2: Recognition metrics for instrumental solo activity.

detection between the *extended* methods. The methods that combine our novel features have a distinct advantage, with the FMRH feature setting achieving the highest precision. Note, however, that the precision ranking of the individual features is different from the vocal case, where the F feature was best and the M and R features showed very similar performance: the method using the R feature alone is now substantially better than that of the simple MFCC feature M, suggesting that using the isolated timbre of the solo melody is a decisive advantage. The F feature alone shows low precision, which is expected because pitch fluctuation is high for vocals as well as instrumental solos.

Considering that the precision of a random classifier in this task is 12.6% the best performance of 61.1%—though not ideal—makes it interesting for practical applications. For example, in a situation where a TV editor requires an expressive instrumental as a musical backdrop to the video footage, a system implementing our method could substantially reduce the amount of time needed to find suitable excerpts.

6. DISCUSSION AND FUTURE WORK

A capability of the *extended* methods we have not discussed in this paper is to detect whether the singer in a song is male or female. A simple classification method is to take the more frequent of the two cases in a track as the track-

wise estimate, resulting in a 70.1% track-wise accuracy. In this context, we are currently investigating hierarchical time series models that allow us to represent a global song model, e.g. 'female song', 'female-male duet' or 'instrumental'. Informal experiments have shown that this strategy can increase overall accuracy, and as a side-effect it delivers a song-level classification which can be used to distinguish not only whether a track's lead vocal is male or female, but also whether the song has vocals at all.

7. CONCLUSIONS

We have proposed the usage of a set of four audio features and the new task of detecting instrumental solos in polyphonic audio recordings of popular music. Among the four proposed audio features three are based on a prior transcription of the predominant melody line, and have not been used in the context of vocal/instrumental activity detection. We conducted 14 different experiments with 7 feature combinations and two different SVM-HMM models. Training and testing was done using 5-fold cross-validation on a set of 102 popular music tracks. Our results demonstrate the benefit of combining the four proposed features. The best performance for vocal detection is achieved by using all four features, leading to a top accuracy of 87.2% and a satisfactory segmentation performance of 72.4%. The detection of instrumental solos equally benefits from the combination of all features. Accuracy is also high (89.8%), but we argue that the main improvement through the features can be seen in the increase in precision to 61.1%. With this paper we also release to the public the annotations we used for training and testing. The annotations offer not only vocal/non-vocal labels, but also distinguish between female and male singers, and different solo instruments.

This work was supported in part by CrestMuse, CREST, JST. Further thanks to Queen Mary University of London and Last.fm for their support.

8. REFERENCES

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, 2003.

[2] A.L. Berenzweig and D.P.W. Ellis. Locating singing voice segments within music signals. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 119–122. IEEE, 2001.

[3] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H.G. Okuno. Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In *8th IEEE International Symposium on Multimedia (ISM'06)*, pages 257–264, 2006.

[4] Masataka Goto. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[5] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC Music Database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 287–288, 2002.

[6] Christopher Harte and Mark Sandler. Automatic chord identifcation using a quantised chromagram. In *Proceedings of 118th Convention*. Audio Engineering Society, 2005.

[7] T. Joachims, T. Finley, and C.N.J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[8] H. Lukashevich, M. Gruhne, and C. Dittmar. Effective singing voice detection in popular music using arma filtering. In *Workshop on Digital Audio Effects (DAFx'07)*, 2007.

[9] N.C. Maddage, K. Wan, C. Xu, and Y. Wang. Singing voice detection using twice-iterated composite fourier transform. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2004)*, volume 2, 2004.

[10] Matthias Mauch. *Automatic Chord Transcription from Audio Using Computational Models of Musical Context*. PhD thesis, Queen Mary University of London, 2010.

[11] P. Mermelstein. Distance measures for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, pages 708–711, 1978.

[12] T.L. Nwe and H. Li. On fusion of timbre-motivated features for singing voice detection and singer identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pages 2225–2228. IEEE, 2008.

[13] Lawrence R. Rabiner and Ronald W. Schafer. *Introduction to Digital Speech Processing*. Now Publishers Inc., 2007.

[14] L. Regnier and G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, 2009.