Vocal Detection in Monaural Mixtures

Anders Elowsson **Ragnar Schön** **Matts Höglund** **Elias Zea** **Anders Friberg**
KTH, Royal Institute of Technology
elov@kth.se rschon@kth.se mattsho@kth.se zea@kth.se afriberg@kth.se

ABSTRACT

In this study, the task of identifying vocals in monaural music mixtures is explored. We show how presently available algorithms for source separation and predominant f0 estimation can be used as a front end from which features can be extracted. A large set of features is presented, devised to connect different vocal cues to the presence of vocals. Two main cues are utilized; the voice is neither stable in pitch nor in timbre. We evaluate the performance of the model by estimating the length of the vocal regions of the mixtures. To facilitate this, a new set of annotations to a widely adopted data set is developed and made available to the community. The proposed model is able to explain about 78 % of the variance in vocal region length. In a classification task, where the excerpts are classified as either vocal or non-vocal, the model has an accuracy of about 0.94.

1. INTRODUCTION

Western music is commonly comprised of regions with singing voice over background music or regions with only background music. As singing voice is an essential element in most western music, the vocal regions should be particularly important to the understanding of the music piece. The detection of these regions can be used as a front end for subsequent processing in many application related to music information retrieval. This strategy has been employed in artist identification [1], query-by-example [2], predominant f0 estimation [3] and popular music clustering [4].

The voice differs from most other instruments in two important regards. It is neither stable in pitch nor in timbre. The fluctuations in pitch are often more rapid and cover a greater range than those produced by e.g. vibrato in other instruments. And although the timbral structure varies for most instruments, the voice has a particularly polytimbral spectrum due to strong and varying formants used to form vowels as well as the varying spectra of unvoiced regions.

Given the important role of the vocal in western music, it is not surprising that singing voice detection has been investigated by several authors. In [5] sinusoidal partials are extracted from the mixtures and each partial's frequency modulation and amplitude modulation are studied to group partials that belong to the voice. In another approach proposed by Ramona et al. [6], an SVM is used on a set of frame based low-level features. The features include spectral descriptors such as spectral centroid and spectral flux (SF) as well as Mel-Frequency Cepstral Coefficients (MFCCs) with derivatives. Also included are features from the monophonic transcription algorithm YIN.

The cues that differentiate the vocal from other instruments have also been used in source separation models applied to monaural mixtures to extract the vocal. A common strategy is to separate the voice implicitly by locating and subtracting sources which are stationary in time or have a broad band frequency spectra. One example of this approach is [7] which is based on a median filtering approach. The underlying idea is that percussive instruments form stable vertical ridges in a spectrogram while harmonic content forms stable vertical lines. By median filtering the spectrogram along time, percussive elements can be detected, whereas by median filtering the signal along the frequency axis, harmonic content can be detected. When the spectrogram has a high frequency resolution, the vocals will be separated with the percussive sources and when a low frequency resolution is used, the vocals will be separated with the harmonic sources. By applying the median filtering twice with different resolutions the vocal can be extracted. In order to further improve the separation performance and to remove some artifacts, FitzGerald and Gainza have also added a tensor factorisation-based separation and a non-negative partial cofactorisation [7].

Other approaches utilize the sparsity of the vocal channel directly. In [8], the *k* nearest neighbors of a spectrogram frame are detected and a soft mask for the background music is created by median filtering these neighbors. In [9], the subspace structures of singing voice and instrumental sounds are learned and the source separation is based on online dictionary learning of these subspaces.

Commonly, the vocal is the strongest f0 in a music mixture when it is present. Therefore, a predominant f0 estimation algorithm could plausibly provide information about the presence of vocals. One such algorithm is *Melodia* [10], which extracts the predominant melody F0 by using a salience based algorithm. First, a sinusoid extraction finds the spectral peaks. A salience function which maps pitch salience over time is constructed by summing the harmonics in each frame. The peaks of this time-frequency representation of the signal are considered potential F0 candidates for the melody. These are then grouped into continuous sequences called pitch contours that may be short single notes or longer phrases. At this stage, non-salient peaks are removed before the contour is characterized as part of the predominant melody or not. This is done by using pitch, length and contour salience based features and their respective distributions.

Vibrato is also used and is shown to be an important feature. Any octave duplicates and pitch outliers are removed. Melodia performed very well in MIREX 2011[1].

In this study we investigate how available algorithms for source separation [7] and predominant f0 detection [10] can be used to identify the presence of singing voice. We calculate features from the output of these algorithms and present their accuracy in voice detection. Finally we devise a model that can be used to estimate the amount of vocals in a musical mixture.

2. DATASET

The mixtures are monaural recordings from the Ballroom dataset, which has been widely used for modelling tempo and beat estimation [11] as well as genre. The dataset was annotated by the authors, with one annotator per song, by aurally identifying the regions with vocals and summing the total length of these regions, rounded to the nearest second. The dataset is comprised of 698 Musical Excerpts (MEs) with a length of 30 seconds each. Figure 1 shows a histogram of the annotated number of seconds of vocal of the MEs in the dataset.

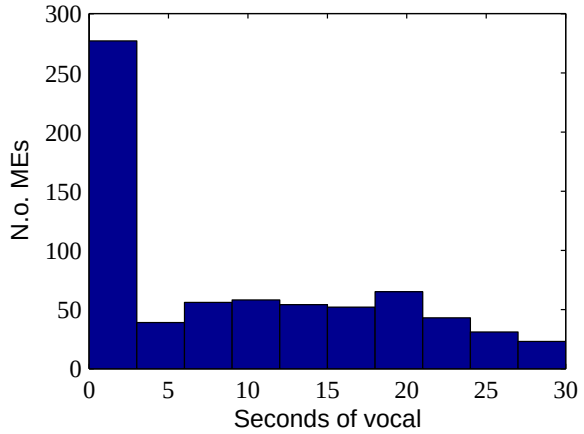

Figure 1. Histogram of the length of the vocal regions for the 698 MEs of the dataset.

A total of 255 MEs contained no vocal and the length of the vocal regions was fairly evenly distributed for the remaining 453 MEs.

3. FEATURES

Having noted that the voice is neither stable in pitch nor in timbre, it seems natural to exploit this phenomena in the feature extraction procedure. Two main approaches were considered in this study.

One approach was to utilize source separation to extract the vocal. The vocal track could then be examined separately, identifying both timbral features as well as the sound level of the vocal track, and comparing this to the timbral features and sound level of the instrumental tracks. This approach was used for the *energy based* features in Section 3.1 and the *timbral* features in Section 3.2.

[1] http://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results

Another approach was to extract the predominant melody of the music and examine the pitch curve of that melody. In order to do this the Melodia melody extraction plugin by Salamon and Gomez [10] was used. This is further explained in Section 3.3.

Furthermore, high level features such as harmonic cues, onset densities and tempo were also included in order to examine the impact on prediction (Section 3.4). These were the features presented in [12] as well as high level features from MIR Toolbox [13].

3.1 Energy-based

The sound level of the vocal track, in relation to the sound level of the instrumental track was used to extract low level features. When using [7] for source separation, three different sources are extracted as the instrumental part consists of a harmonic track (containing instruments with steady partials) and a percussive track. Therefore the energy of the vocal track was compared with both these tracks separately. The frequency spectra was divided into five frequency bands, each covering approximately two octaves of the spectrum, with center frequencies 70 Hz, 250 Hz, 600 Hz, 1.1 kHz and 3.3 kHz.

By dividing the sound into frames of one second each, different measures of the difference in energy between the two different sources were defined. These were the mean difference, the difference between the average of the seven highest energy frames for each source and the difference between the median energy frame for each source. The motivation behind the ordinal features was that they could be used to pick up how loud the vocal track was mixed in the music. Assuming that the vocal track is mixed at different volumes for different songs, the difference in energy between the vocal track and the instrumental track becomes a crude measurement. This was also taken into account with a separate set of features by tracking the sound level of the vocal track over time and comparing it with the sound level of the instrumental track. The vocal was defined as present if the low pass filtered vocal track was within 16 dB of the low pass filtered instrumental track.

3.2 Timbral

It is plausible that timbral features of the vocal track can be used to identify the presence of singing. The polytimbral spectrum of the voice, due to strong and varying formants, is an essential cue. For this category, a large number of features related to pitch, dynamics, tonality, timbre and rhythm were thus calculated using the MIR Toolbox [13]. Among these features are the MFCCs, which is a popular means for describing timbre. MFCCs are calculated by taking a discrete cosinus transform of the Mel-scaled spectrum. This can be interpreted as the spectral shape of the timbre.

In this study, the standard 39 MFCC's were calculated, including the first and second derivatives (called delta and delta-delta respectively). Other features in this category include chromagram, key strength, spectral centroid, flux, flatness, entropy, rolloff, brightness and roughness

among others. Together, these features provide a comprehensive description of the audio in question. However, many of these are local features and to match the scope of this study they were transformed into global features by calculating their mean and variance across the ME.

3.3 Pitch-based

The pitch-based features were designed with the predominant f0 estimation algorithm *Melodia* [10] as a front end. The algorithm was applied to the non-separated mixtures and the output was post-processed before the features were computed. First, negative values were removed, as they are generated when Melodia is unable to determine any predominant f0. Also, all values above 400 Hz were removed as the fundamental frequency of the voice of adult humans is typically below 400 Hz. At this point, a number of general features such as the mean, median and standard deviation of the predominant f0 was extracted. For the remaining features, fluctuations in pitch larger than 50 cent between two adjacent frames were ignored. This was done to avoid any influence of octave errors or pitch shifts from new notes. The resulting array was median filtered in time with a kernel size that varied between 8.9 and 66.7 ms (3-23 frames) and the absolute value of the difference of the median filtered array was computed and summed for each frame *n* as in Equation 1.

$$shift(n) = \sum_{i=n}^{n+15} \left| F_0(i) - F_0(i-k) \right| \qquad (1)$$

By varying *k* and the length of the median filter, and by also testing different limits which restricted too small shifts from being included, different features could be extracted. Note that the pitch shifts of 15 consecutive frames are summed in the *shift* array.

3.4 High level

An onset density function was calculated in two different ways. A rhythmic onset density was based on the SF of the source separated percussive track. A harmonic onset density was extracted from the SF of the Contant-Q Transform (CQT) of the original waveform as described in [14]. The algorithm uses a vibrato suppression scheme by subtracting the sound level of each bin of the new frame with the maximum sound level of the adjacent bins in the old frame. In this way, shifts of a peak by 20 cents (one bin in the CQT) are restricted from affecting the SF.

4. MODELS

The modelling consisted of three main parts; Feature selection, linear regression and classification, in order to measure the different features ability to distinguish between vocal or non-vocal content.

4.1 Training and test set

The ballroom dataset was divided into a training set and a test set. From the 698 MEs in the dataset, 20 % (140 MEs) were randomly selected for the test set and the remaining 80 % (558 MEs) composed the training set. The test set was kept hidden, and only used to measure the accuracy of the final regression model and classification.

4.2 Linear regression and feature selection

A linear regression model was fitted to the data in order to both estimate the quality of the different groups of features and select a subset of features from the total feature space. Due to the high dimension of the feature space, the dimensionality was reduced. First, linearly dependent features were removed from the data set, and then a forward sequential feature selection was performed. This feature selection method first chooses the feature with the lowest mean squared error (MSE), computed from the linear regression. The chosen feature is added to a reduced feature set. Next, features are iteratively added to the reduced feature set by choosing the feature that together with the already selected features estimates the duration with the lowest error. For each evaluated feature, the error measurements are cross validated (10-fold) on the training set. The procedure is continued until the performance converges. It was repeated for each group of features and was also applied to a merged group of all features. The computed feature coefficients from the training set were finally applied to estimate the duration of the vocal regions in the test set.

4.3 Classification

After estimating the durations of the vocal regions, classification was introduced. The classification task consisted of classifying the MEs into two groups, one for the vocal and one for the instrumental MEs. As songs with just a few seconds of vocals would not fit well into either group they were simply removed from this task. Setting the limit for an ME to be considered vocal to a minimum of 8 seconds the two groups consisted of 255 MES (instrumental) and 360 MEs (vocal) respectively. Two different classification methods were used.

The first approach was to use linear discriminant analysis (LDA). Assuming that non-vocal and vocal content are both normally distributed with different means but the same variances, the classifier finds a linear threshold that divides the different clusters in two. The test sample was here classified as belonging to the cluster to which is has the highest probability to belong to according to the normal distributed means and variance.

The second approach was to use logistic regression, which is a probabilistic classification model. In Logistic regression the logistic function is used to map the range of negative infinity to positive infinity into the range of 0 to 1. This time we further refined the feature space by selecting only the best explanatory features given from the linear regression. This selection step was also carried out with forward selection and a maximum of 22 features were used.

5. RESULTS

5.1 Correlation of individual features

In Sections 5.1.1-5.1.4 the features with the highest correlations are presented, for each group.

5.1.1 Energy-based

The best explaining features based on vocal energy had a correlation of about 0.63. They were taken from the difference in energy of the separated vocal track and the harmonic track. The frequency bands with center frequency 600 Hz and 1.1 kHz had the highest correlation.

5.1.2 Timbral

Among the timbral features, the logarithm of the variance of the MIR Toolbox feature *spectral brightness* computed from the vocal track was the most important, reaching a correlation of about 0.49.

5.1.3 Pitch-based

The best explaining feature based on the pitch of the predominant melody was related to the change in pitch over time and had a correlation of 0.67. The specific feature was computed as described in Section 3.3, with a median filter length of 7 frames, a window size of 15 frames and a limit of 100 cents.

5.1.4 High level features

The Harmonic Change Detection Function (HCDF) from the MIR Toolbox was the best high level feature. The logarithm of the feature, computed as the mean HCDF of the whole ME was used.

5.2 Regression

Table 1 shows the accuracy in predicting the total length of the regions of vocals in the MEs. When using all features 78 % of the variance could be explained.

Group N.o.	**1**	**2**	**3**	**4**	**All**
R^2	0.60	0.68	0.59	0.38	0.78
MSE	0.59	0.44	0.42	0.75	0.36

Table 1. The result of the linear regression, with R^2 representing the squared correlation and MSE the mean squared error of the regression.

In Figure 2 the predicted length of vocals in the MEs are compared with the annotated length. The data for the Figure is based on a multiple linear regression run on all MEs with selected features.

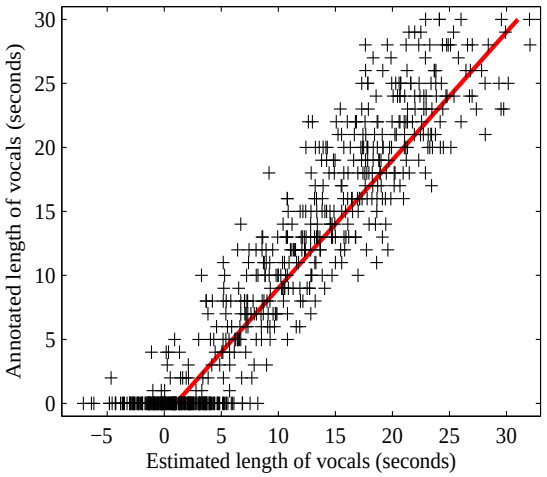

Figure 2. The estimated length of vocal regions in relation to the annotated length. Each ME is marked with a cross.

5.3 Classification

In Table 2, the results of the classification task are presented. Evidently the logistic regression performed somewhat better than LDA for this task. We note that all groups of features are relatively good at classifying the MEs.

Group No.	**1**	**2**	**3**	**4**	**All**
LDA Precision	0.77	0.85	0.75	0.89	0.82
LDA Recall	0.76	0.85	0.95	0.68	0.91
LDA Accuracy	**0.72**	**0.82**	**0.82**	**0.69**	**0.84**
LR Precision	0.93	0.93	0.94	0.89	0.95
LR Recall	0.92	0.93	0.91	0.90	0.93
LR Accuracy	**0.91**	**0.92**	**0.91**	**0.88**	**0.94**

Table 2. The result of the classification using two different classification methods LDA and logistic regression (LR).

6. CONCLUSIONS AND DISCUSSION

A set of features has been proposed and combined into a model which can be used to identify vocals in monaural music. The features are based on two cues which distinguish the voice from other instruments; rapid shifts in pitch and timbre. By utilizing presently available algorithms for source separation (predominant f0 estimation and MIR-related feature extraction algorithms), the presented work should be feasible to replicate to other researcher in the field. The proposed model is able to explain about 78 % of the variance in vocal region length. In a classification task, where the excerpts are classified as either vocal or non-vocal, the model has an accuracy of about 0.94.

FitzGerald and Gainza [7] suggest that performance of their source separation algorithm could be further

improved by distinguishing between vocal and non-vocal regions. It is our hope that this study can contribute to that endeavor. We note that even the best energy-based features can only explain about 40 % of the variance in vocal region length (correlation of 0.63 gives an R^2 of 0.4); unlike the full model which is able to explain 78 % of the variance. The implication is that attempts to discern between vocal and non-vocal regions could benefit from a model which uses several cues to distinguish vocals from other musical instruments.

The features were divided into different groups to show how well different aspects can explain the presence of vocals. We found the highest correlations in the individual features that were based on the difference in energy between the separated vocal and the harmonic track, as well as pitch shifts in the predominant melody. For some features it was difficult to find the most suitable group. As an example the high-level feature HCDF could arguably be included in both the pitch group and the timbre group.

The annotations from this study are freely available for research purposes[2].

Acknowledgments

This study was initiated during the SMC summer school of 2013. We wish to thank our fellow group member Hama Biglari for his valuable contribution. We wish to thank Derry FitzGerald for kindly providing source code.

7. REFERENCES

[1] A. Berenzweig, D. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proc. AES 22nd Int. Conf. Virtual, Synthetic Entertainment Audio*, 2002.

[2] W. H. Tsai, H. M. Yu, and H. M. Wang. Query-by-example technique for retrieving cover versions of popular songs with similar melodies. *In Proc. of ISMIR'05*, 2005.

[3] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H.G. Okuno, "F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. V-253-V-256, Toulouse, 14-19, May, 2006.

[4] W.-H. Tsai, D. Rogers, and H.-M. Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music J.*, vol. 28, no. 3, pp. 68–78, 2004.

[5] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," *IEEE ICASSP*, pp. 1685-1688, 2009.

[6] M. Ramona, G. Richard, and B. David. "Vocal detection in music with support vector machines," in *IEEE International Conference on Acoust. Speech and Sig. Process*, 2008.

[7] D. FitzGerald and M. Gainza, "Single Channel Vocal Separation using Median Filtering and Factorisation Techniques", ISAST Transactions on Electronic and Signal Processing , No. 1, Vol. 4,2010 (ISSN 1797-2329), pages: 62 - 73, 2010.

[8] D. FitzGerald, "Vocal Separation Using Nearest Neighbours and Median Filtering", in *23nd IET Irish Signals and Systems Conference*, NUI Maynooth, 2012.

[9] Y. H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries", *In Proc. of ISMIR*, 2013.

[10] J. Salamon and E. Gómez, "Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics", IEEE Transactions on Audio, Speech and Language Processing, 20(6):1759-1770, Aug. 2012.

[11] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. "An experimental comparison of audio tempo induction algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 5, 1832-1844, 2006.

[12] A. Elowsson, and A. Friberg, "Modelling the speed of music using features from harmonic/percussive separated audio," in *Proc. of ISMIR*, 481-486, 2013.

[13] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction From Audio", *International Conference on Digital Audio Effects*, Bordeaux, 2007.

[14] A. Elowsson, and A. Friberg, "Modelling perception of speed in music audio," in *Proc. of the Sound and Music Computing Conference 2013*," 735-741, 2013.

[2] Made available for research purposes at www.speech.kth.se/music/voice