Singing Voice Detection for Karaoke Application

Arun Shenoy[*], Yuansheng Wu[†], Ye Wang[†]

ABSTRACT

We present a framework to detect the regions of singing voice in musical audio signals. This work is oriented towards the development of a robust transcriber of lyrics for karaoke applications. The technique leverages on a combination of low-level audio features and higher level musical knowledge of rhythm and tonality. Musical knowledge of the key is used to create a song-specific filterbank to attenuate the presence of the pitched musical instruments. This is followed by subband processing of the audio to detect the musical octaves in which the vocals are present. Text processing is employed to approximate the duration of the sung passages using freely available lyrics. This is used to obtain a dynamic threshold for vocal/ non-vocal segmentation. This pairing of audio and text processing helps create a more accurate system. Experimental evaluation on a small database of popular songs shows the validity of the proposed approach. Holistic and per-component evaluation of the system is conducted and various improvements are discussed.

Keywords: Karaoke, **s**inging voice, vocal segmentation, tonic, key, inverse comb filtering, rhythm, lyrics.

1. INTRODUCTION

Karaoke is a Japanese abbreviated compound word, "kara" comes from "karappo" meaning empty, and "oke" is the abbreviation of "okesutura," or orchestra. Usually, a recorded popular song consists of vocals and accompaniment. Musical works in which only the accompaniment is recorded were named "karaoke." Karaoke singing involves singing to such recorded accompaniments of popular songs in front of a live audience. After the singer chooses a song from a catalogue, lyrics are usually displayed on a monitor, recorded music plays, and it's showtime for the novice pop star. Invented in the late 1970's, the wild popularity of karaoke over the years has swept this form of singing into the mainstream throughout the world. Karaoke creates its own culture, while reflecting much about the wider culture and the place of popular music as a media form[6].

It would be commercially very useful to develop a computational karaoke model that could analyze a musical recording and transcribe the lyrics, but this is currently impractical. Transcription of lyrics using speech recognition is an extremely challenging task as singing differs from speech in many ways. The phonetic and timing modification, presence of meaningless syllables often employed by singers and interference of the instrumental background would make an acoustic classifier trained on normal speech a poor match to the acoustics of the sung vocal line.

This difficulty has led us to re-examine the transcription problem[31]. We recognize that transcription is often not necessary, as many lyrics are already freely available on the Internet. However, text based lyrics do not provide any timing information. Thus, the main task involved in the process of karaoke is embedding lyrical time stamps inside the musical audio file. This kind of an alignment is currently a manual process. Towards this end, we have developed a prototype[31] to automate this process of forced alignment between the music and the lyrics, saving manual labor. One of the key components of this framework is singing voice detection, a precursor, for this sort of forced alignment. The approach to this problem[31], employs the use of stochastic classifier that modeled musically relevant song structure information in addition to traditional audio features. In the current work, we propose a simpler rule-based approach to this problem that leverages on the combination of low-level audio features and higher level music knowledge of rhythm and tonality.

[*] [†] School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543
[*] arun@arunshenoy.com, [†]{wuyuansh,wangye}@comp.nus.edu.sg

Visual Communications and Image Processing 2005, edited by Shipeng Li, Fernando Pereira, Heung-Yeung Shum, Andrew G. Tescher, Proc. of SPIE Vol. 5960 (SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.631645

2. RELATED WORK

The singing voice, in addition to being the oldest musical instrument, is also one of the most complex from an acoustic standpoint[11]. Research on the perception of singing is not as developed as in the closely related field of speech research[26]. Some of the existing work is surveyed in this section.

Chou and Gu[5] have utilized a gaussian mixture model (GMM) to detect the vocal regions. The feature vectors used for the GMM include 4Hz modulation energy, harmonic coefficients, 4 Hz harmonic coefficients, delta mel frequency cepstral coefficients (MFCC) and delta log energy.

Berenzweig and Ellis[3] have used a speech recognizer's classifier to distinguish vocal segments from accompaniment. It has been mentioned that though singing is quite different from normal speech, it shares some attributes of regular speech such as formant structure and phone transitions. Thus a speech-trained acoustic model might respond in a detectably different manner to singing than to other instruments. Three broad feature sets have been explored, basic posterior probability features (PPFs), derived statistics such as classifier entropy and average of these values. Within music, the resemblance between the singing voice and natural speech will tend to shift the behavior of the PPFs closer towards the characteristics of natural speech when compared to non-vocal instrumentation.

Berenzweig et al.[4] have proposed a technique to improve artist classification of music using voice segments. The basic paradigm of the system was to classify musical tracks as being the work of one of several predefined artists. This is a two stage process comprising of vocal segmentation using a two-class multi-layer perceptron (MLP) neural net trained with hand-labeled data followed by artist classification also performed by an MLP neural network. For the purpose of the current work, we shall focus only on the singing voice detection schemes discussed in the literature. The features used for the vocal segmentation task comprised of 13 PLP coefficients along with deltas and double deltas. To segment the data, the PLP features are calculated and fed to the segmentation network. The output is a stream of posterior probabilities of the two classes (vocal and instrumental music) which is compared against a threshold. It has been highlighted that this approach is far simpler as compared to the earlier one[3]. This is however sufficient for the purpose of artist classification as all the vocal segments need not be identified, just a sufficient percentage with a low error rate.

Kim and Whitman[11] have developed a system for singer identification in popular music recordings using voice coding features. As a first step, an untrained algorithm is used to automatically extract vocal segments. Once these segments are identified, they are presented to a trained singer identification system. To detect the singing voice, the audio signal is first filtered with a band-pass filter which allows the vocal range (200-2000 Hz) to pass through while attenuating other frequency regions. This is achieved via a simple chebychev infinite-impulse response (IIR) digital filter. To further filter out other instruments producing energy in this region (like the drums), an inverse comb filterbank is then applied to obtain the fundamental frequency at which the signal is most attenuated. The harmonicity has been defined as the ratio of the total signal energy to the maximally harmonically attenuated signal. By thresholding the harmonicity against a fixed value, a detector for harmonic sounds is obtained. The hypothesis is that most of these correspond to regions of singing voice based on its highly harmonic nature when compared to other high energy sounds in the vocal band.

Another system for automatic singer identification has been proposed by Zhang[32]. This is a two step process comprising of a training phase, during which a statistical model is created for a singer's voice, and a working phase, during which the starting point of the singing voice is detected and a fixed length of testing data is taken from that point. Audio features extracted from this data are then compared against the existing singers' models to perform singer identification. Singing voice detection is achieved by extracting features of energy, average zero-crossing rate (ZCR), harmonic coefficients and spectral flux computed at regular intervals which are then compared against a set of predetermined thresholds.

A system for the blind clustering of popular music recordings based on singer voice characteristics has been proposed by Tsai et al.[28]. Methods are presented to separate vocal and non-vocal regions, model singers' vocal characteristics and clustering of recordings based on singer characteristic similarity. The singing voice detection is done in two stages. In the first stage, the training phase, a statistical classifier with parametric models is trained using the manual vocal/non-vocal transcriptions of the singer's voice. Two separate GMMs are used for this task, a vocal GMM and a non-vocal GMM. In

the testing phase, the recognizer takes as input the feature vector extracted from an unknown recording and produces as output, the likelihood for the vocal and non-vocal GMM. The feature vector used in the system was the MFCC.

A system for automatic detection and tracking of target singer in multi-singer recordings has been presented by Tsai and Wang[29]. Methods are presented to separate vocal and non-vocal regions, model singers' vocal characteristics and to distinguish a target singer from other simultaneous or non-simultaneous singers. The vocal and non-vocal classification has been achieved using a stochastic classifier that consists of a front-end signal processor to extract cepstrum-based feature vectors, followed by a backend statistical processor that performs modeling and matching. It operates in 2 phases, training and testing. In the training phase, a music database with manually annotated vocal and non-vocal regions is used to create a set of three GMMs to characterize vocal and non-vocal classes. The first GMM is formed using the labeled vocals regions of a target singer. The second and third one are trained using the manually annotated vocal and non-vocal regions of all the music data available. During testing, the classifier takes as input a feature vector extracted from an unknown recording and calculates the likelihood to the trained GMMs.

Bartsch[2] has proposed a system for automatic singer identification in popular music. A separation system known as PESCE has been designed to achieve two separate goals, singing voice detection and singing voice extraction. This system is effectively a fundamental frequency estimation algorithm for polyphonic music. It takes a short audio signal as input, and it produces fundamental frequency estimates of voice-like sources that are present in the signal. PESCE assumes that the partials of the singing voice have significant frequency modulation while other instruments have constant-frequency partials. Thus, voice-like sources are those that exhibit significant frequency modulation. If no voice-like sources are present, PESCE will produce no output. The fundamental frequency estimate will allow one to extract time-varying amplitudes for the partials of the voice signal from a time-frequency distribution such as the spectrogram. This extraction has been referred to as separating the voice signal, since the singing voice partials are being separated from partials that arise from other instruments.

Nwe and Wang[19] have proposed a statistical model to classify segments of musical audio into vocal or non-vocal using a Hidden Markov Model (HMM) classifier. The feature extraction is based on sub-band processing that uses the log frequency power coefficients (LFPC) to provide an indication of the energy distribution among subbands. The training model also takes into account tempo and song structure information in song modeling based on the observed variations in intra-song signal characteristics. Thus, in contrast to conventional HMM training methods that employ one model for each class, the method here uses a multi-model HMM technique to allow for more accurate modeling as compared to the single-model baseline. A bootstrapped HMM has been used to further increase the classification accuracy.

Nwe et al.[20] have enhanced the previously discussed model to incorporate musically relevant quarter-note spaced segmentation followed by harmonic attenuation of the input signal using the frequencies in the key of the song.

Maddage(a) et al.[13] have adopted a twice-iterated composite fourier transform (TICFT) technique to detect the singing voice boundaries. The TICFT is computed over each frame where the magnitude spectrum of the first FT is input to a second FFT. Singing voice frames are separated from instrumental frames based on a linear threshold set on the energy of the second FFT spectrum. A statistical autocorrelation of the bass and snare drum onset times is used to frame the audio into quarter-note spaced segments. Heuristic rules based on musical chord change patterns have been extended to apply to the singing voice to further increase the accuracy of vocal detection.

Maddage(b) et al.[14] have proposed a technique to detect semantic regions in musical audio using support vector machines (SVM) and GMMs as classifiers. A statistical autocorrelation of the bass and snare drum onset times is used to frame the audio into quarter-note spaced segments. The audio feature used is the Cepstral coefficients extracted from the musically based octave-scaled subbands as well as from the perceptually based mel-scaled subbands. Singular value decomposition has been applied in both cases to find the uncorrelated Cepstral coefficients. Experimental results have shown that the SVM performs better than GMM and that the octave-scaling performs better that the mel-scaling of the audio for feature extraction.
Maddage(c) et al.[15] have proposed a framework for music structure analysis with the help of repeated chord pattern analysis and vocal content analysis. The vocal boundary detection in this work is similar to the one proposed earlier[14]. Only SVM has been used as the classifier and heuristic rules based on the rhythm structure of the song have been applied

to further increase the accuracy of vocal detection. The same technique has been used by Maddage(d) et al.[16] in a singer identification framework based on vocal and instrument models.

Tzanetakis[30] has proposed a semi-automatic approach to the problem of locating singing voice segments. In this approach, a small random sampling of the song is manually annotated and the information learned is used to automatically infer the singing voice structure of the entire song. Thus a different classifier is trained for each song using the bootstrapping annotation information for training. The feature set used consists of the following: mean and standard deviation of the centroid, rolloff and flux and the mean relative energy of the subbands that spans the lowest ¼ and the second ¼ of the total bandwidth. In addition the mean and standard deviation of the pitch were also used. A wide range of classifiers were used to compare performance in the bootstrapping and classification task. The best generalization performance was obtained using the logical regression classifier and the neural network.

3. SYSTEM DESCRIPTION

Our framework comprises of five stages as shown in Figure 1. Each stage will utilize the information derived from the previous stage.

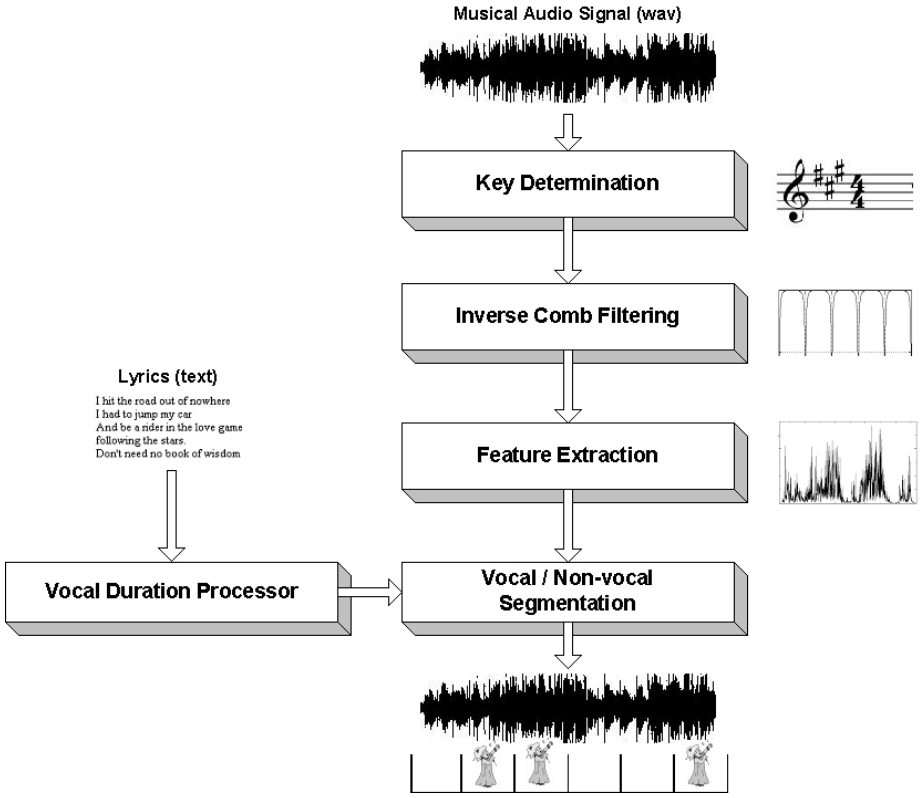

Figure 1: System description

3.1 Key determination

Rhythm is a component that is fundamental to the perception of music. It can be perceived as a combination of strong and weak beats. A strong beat usually corresponds to the first and third quarter note in a measure and the weak beat corresponds to the second and fourth quarter note in a measure[7]. If the strong beat constantly alternates with the weak beat, the inter-beat-interval (the temporal difference between two successive beats), would correspond to the temporal length of a *quarter note*. The audio has been framed into beat-length segments to extract metadata in the form of quarter note detection of the music. The basis for this technique is to assist in the detection of chord structure and subsequently

the *key* [25], based on the musical knowledge that chords are more likely to change at beat times than on other positions[8]. The knowledge of the musical key will serve as an input to the next stage.

3.2 Inverse Comb Filtering

Tonic is sometimes used interchangeably with key. The word tonic simply refers to the most important note in a piece or section of a piece. Music that follows this principle is called tonal music. In the tonal system, all the notes are perceived in relation to one central or stable pitch, the tonic. All tonal music is based upon scales. The tonic/key defines the diatonic scale which a piece of music uses (most familiar as the Major/Minor scale in music).

We run the beat spaced audio frames through a series of inverse comb filters which attenuate the signal at the frequencies (and the corresponding harmonics) in the key of the song. This would serve to remove the presence of the pitched instruments. This is shown in Figure 2 below.

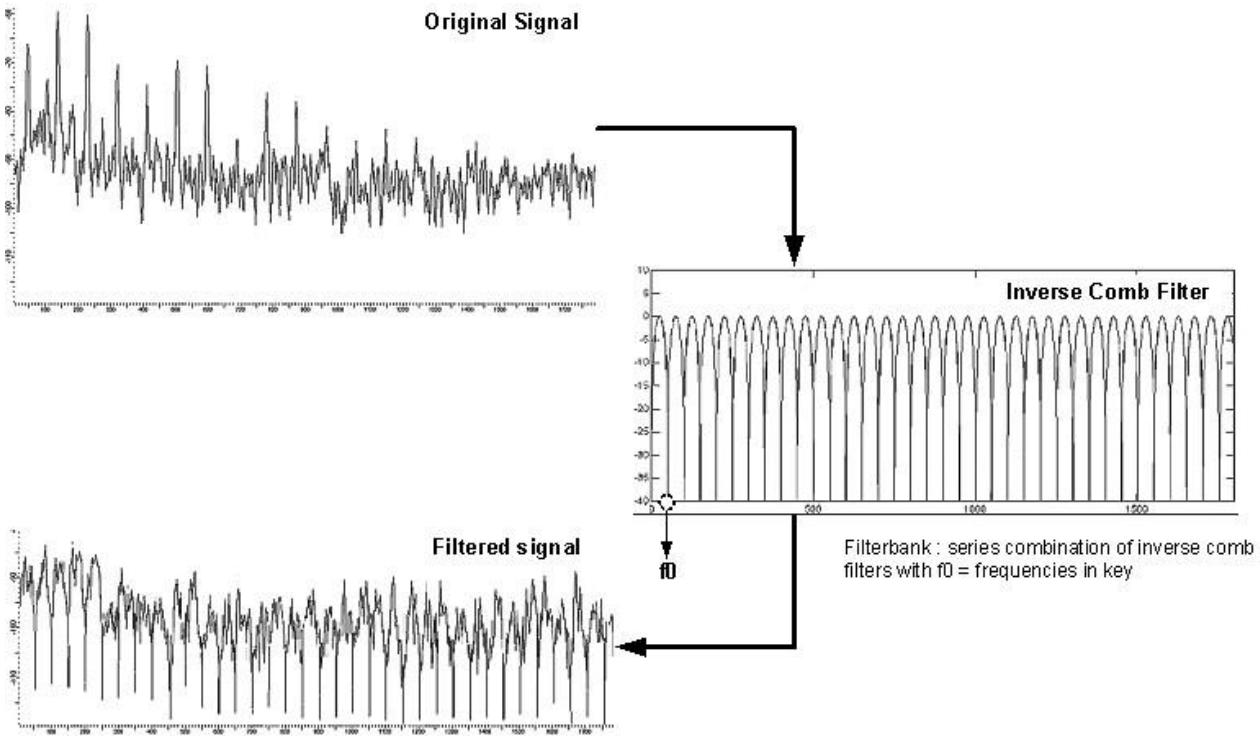

Figure 2: Key Filtering

An interesting observation is that though the singing voice falls under the category of pitched musical instruments, it is attenuated only partially as compared to the other pitched musical instruments. At the onset, this would appear rather strange, because the singing voice is more than 90% *voiced*[11]. Singing primarily consists of sounds generated by phonation, the rapid vibration of the vocal folds resulting in utterances referred to as voiced. This is as opposed to *unvoiced* sounds which are generated by the turbulence of air against the lips or tongue such as the consonants ‘f’ or ‘s’. Because of the harmonic nature of voiced speech, the majority of the energy would resides in its harmonics[18] and hence, theoretically speaking, be removed by the inverse comb filter. But the fact that it is not, can be attributed to two important aspects of singers' F0 control: vibrato and intonation.

3.2.1 Vibrato

From an acoustic perspective, the vibrato is defined as a regular fluctuation in the pitch of the signal. It is frequently assumed that the vibrato is useful in musical practice because it reduces the demands on accuracy of fundamental frequency[26]. It is described by two parameters:

- Rate of vibrato: the number of undulations occurring during one second
- Extent of vibrato: depth of the modulation expressed in a percentage of the average frequency. More often, this is expressed in *cent*s, the interval between two tones having the frequency ratio of $1:2^{1/1200}$. An equally tempered semitone is equal to 100 cents.

Seashore[24] has reported that the mean vibrato rate for 29 singers is 6.6 undulations per second and average extent is ± 48 cents. This information could have been used to select a more optimized quality factor (Q factor) q for the filter. However, this is not practical because of two other problems:

- The vibrato rate, though constant for any given singer, varies slightly between singers[24].
- There is a significant vibrato extent in professional western lyric singing for individual tones[21]. The mean vibrato extent for individual tones ranges between ±34 and ±123 cent.

The scope of this work does not include singer identification nor any form of note level transcription. Hence the vibrato cannot be modeled. It should be noted that musical instruments also exhibit a considerable bit of vibrato. However, it has been observed that vibrato extent is lower in musical instruments (0.2-0.35 semitones) as compared to singers (0.6 to 2.0 semitones)[27]. Thus, on key filtering, the attenuation of the musical instruments will be greater than that of the singing voice.

3.2.2 Intonation

Intonation refers to the manner of producing or uttering tones, especially with regard to accuracy of pitch and the exactitude of the pitch relations. The singing voice follows the key of the music and singers modify vocal cord tension to change the pitch to produce the desired musical note.

Two observations have been highlighted by Sundberg[26]:

- The long notes begin slightly flat (about 90 cents on the average), and are gradually corrected during the initial 200 ms of the tone. Moreover, many of these notes change their average frequency in various ways during the course of the tone.
- For short tones, it has been observed that the average fundamental frequency in a *coloratura* (a soprano who sings elaborate ornamentation) passage does not change stepwise between the target frequencies corresponding to the pitches we perceive. Rather, the average rises and falls monotonically at an approximately constant rate. Moreover, difficulties seem to occur when the pitch is very high. In this case, the pitch changes between the scale tones are wide in terms of absolute frequency.

Further, as with vibrato, Prame[21] has noted that intonation substantially departs from equally tempered tuning for individual tones. Deviations from theoretically correct frequencies are used as a means of musical expression. Thus, though the passage is perceived as rapid sequences of discrete pitches, the fundamental frequency events do not form a pattern of discrete fundamental frequencies. This would compromise the accuracy of our computational frequency analysis model. Three other characteristics observed by Saitou et al.[23] should also be considered:

- Overshoot: Deflection exceeding the target note after note changes.
- Preparation: Deflection of the opposite direction of note change observed just before note changes
- Fine-fluctuation: Irregularly fine fluctuation higher than 10 Hz.

We infer that the first two of these can probably be closely correlated with the observations of long and short notes discussed above. As with the vibrato, all the aspects discussed in this section are too complex to incorporate into the current model and hence are not handled by the key filtering technique.

The residual signal, after applying the key filters, would contains a significant presence of the sung vocals in addition to drums (and other unpitched percussive instruments). Most of the pitched instrument presence would be removed.

Harmonic attenuation of the input signal using the frequencies in the key of the song has been incorporated in an earlier work[20]. However the implementation was done with a filterbank of triangular filters spaced on a linear-logarithm scale. This spacing of filters follows the mel frequency scale, which is inspired by critical band measurements of the human auditory system. It has also been used in other work[14-16] that utilize cepstral features derived from the power spectrum. In the current work, the key filtering is implemented using an inverse comb filterbank that attenuates the frequencies in the key of the song and all partials while allowing the rest to pass through. The advantages of this approach are discussed later in this paper. The inverse comb filterbank has been used earlier to find the fundamental frequency at which the signal is most attenuated[11]. This was achieved by using a bank of inverse comb filters with various delays. In contrast, our implementation is more musically motivated, where the frequencies are known apriori.

3.3 Feature Extraction

The acoustic signal can now be perceived to contain the singing voice which has most of its frequency components located around the key frequencies and the percussive sounds which have their frequency components spread more uniformly over the entire frequency region with no prominent frequency spectrum peaks. We now perform sub-band processing of the audio, where each subband spans one Octave in the tempered scale[1]. The majority of the singing voice falls between 200 Hz and 2000 Hz[11]. Hence we consider only the four Octaves that fall in this range, C3 (~130 Hz) to B6 (~1975 Hz). Each quarter-note spaced segment of audio is further segmented into 10 ms frame segments for finer resolution. The signal is assumed to be quasi-stationary during this period. The energy function for each subband is obtained which represents the amplitude variation over time of the musical audio signal[22].

3.4 Vocal Duration Processor

To identify the frames containing vocals, a static threshold cannot be applied as the proportion of the song containing sung vocals varies across songs. Thus, a multi-modal audio-text approach is employed to determine an adaptive threshold based on the duration of the vocals in the song. We have presented a technique to determine the duration of the vocals in the song using only its corresponding textual lyrics[31]. To accomplish this, each word in the lyrics has been first decomposed into its phonemes based on the word's transcription in an inventory of 39 phonemes from the CMU Pronouncing Dictionary. As phoneme durations in sung vocals and speech differ, information from speech recognizers or synthesizers is not used. Rather, a separate database containing around 500 lines of lyrics with manually annotated timing information is used to learn the duration of phonemes. Each line in this sung training database is decomposed into its phonemes and the manually annotated line duration is distributed uniformly among its phonemes. In this way, a phoneme can be modeled by the distribution of its instances. For simplicity, phoneme duration distribution has been modeled as gaussian, characterized by mean and variance. To calculate the vocal duration of the test song, the gaussian distributions representing all phonemes present has been used.

3.5 Vocal/ Non-vocal Segmentation

Vocal frames are normally reflected by a rise in the energy level of the audio. Thus the frames with the highest energy are classified as vocal frames. The number of these frames is selected by a threshold, set adaptively such that the proportion of the frames chosen is equivalent to the proportion of the vocal duration in the entire song as determined by the vocal duration processor.

4. SYSTEM EVALUATION

Our experiments are performed on a database of 10 popular English songs carefully selected for their variety in artist and time spans. We assume the meter to be 4/4, this being the most frequent meter of popular songs and the tempo of the input song is assumed to be constrained between 40-185 M.M. (Mälzels Metronome: the number of quarter notes per minute) and almost constant. The relatively small size of this database is because of the tedious and somewhat ill-defined nature of the task of obtaining ground truth data[11]. Establishing exactly where a vocal segment begins and ends is problematic. Low-level background vocals that tend to fade in out in some songs add further complication. Every effort has been made to keep the segmentation on this set as accurate as possible.

4.1 Results

The holistic and per-component evaluation of the system is presented in Tables 1 and 2 using the traditional measures of retrieval performance, *Recall* (completeness of retrieval) and *Precision* (purity of retrieval). Recall is the ratio of the number of correct vocal frames detected to the total number of hand labeled vocal frames, expressed as a percentage. Precision on the other hand, is used to determine, of the automatically detected frames, how many are correct. This again is expressed as a percentage. By comparison with hand labeled data, we conclude from Table 1 that the overall Recall and Precision rates for the system are 89.44 % and 77.37 % respectively. For a given song, the 2 adjacent subbands that give the highest averaged combination of Precision and Recall have been used to obtain the final result. This is based on the premise that singers possess a dynamic pitch range of 2-Octaves[10] and hence this would reflect the true regions of singing voice.

Table 1: System evaluation

Song	Title	Band 1 Octave 3 (C3 - B3)		Band 2 Octave 4 (C4 - B4)		Band 3 Octave 5 (C5 - B5)		Band 4 Octave 6 (C6 - B6)		Final Accuracy (over 2 bands)	
		Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)
1	Corrs - Breathless	94.36	74.77	93.73	77.34	91.85	77.34	87.77	69.86	92.79	77.34
2	Michael Jackson - Heal the world	99.18	86.90	99.45	87.38	99.45	86.90	98.90	86.43	99.45	87.14
3	Craig David - Walking away	99.10	81.25	99.10	81.62	99.10	81.62	99.10	81.62	99.10	81.62
4	Natalie Imbruglia - Torn	100.00	78.39	100.00	77.86	98.93	75.78	99.29	75.26	100.00	78.13
5	Tina Arena - Burn	82.35	78.98	75.82	72.51	71.9	66.85	69.61	68.46	79.09	75.75
6	Spice Girls - Viva Forever	78.95	74.74	72.18	71.94	70.68	71.68	65.04	65.31	75.57	73.34
7	Westlife - Seasons in the sun	99.22	75.89	99.22	75.89	99.22	75.89	99.22	75.89	99.22	75.89
8	Hanson - Mmmbop	91.89	85.86	89.46	85.16	86.49	83.94	83.51	81.27	90.68	85.51
9	Joan Osborne - One of us	80.95	67.29	83.33	72.39	80.56	69.71	74.21	62.73	81.95	71.05
10	Roxette - It must have been love	68.29	53.92	75.12	66.27	78.05	69.58	72.20	68.07	76.59	67.93
									Net Recall and Precision	**89.44**	**77.37**

The per-component evaluation is presented in Table 2. Errors in key determination do not affect the filtering process. This is explained in more detail in the following section. When compared to the results in Table 1, it is observed that the overall recall and precision drops by 0.72 % and 1.39 % respectively when the filterbank is removed from the framework. Errors in the text duration estimation account for a drop in performance of 2.23 % and 3.39 % for the recall and precision respectively. This is obtained by replacing the vocal duration processor by a manually encoded duration value. The + / - for text duration error in Table 2 represent offsets (expressed as a percentage) from the actual manually calculated duration.

Table 2: Per-component evaluation

Song	Actual Key	Detected Key	Final Accuracy (without key filtering)		Text Duration Error	Final Accuracy (without text processor)	
			Recall (%)	Precision (%)	(%)	Recall (%)	Precision (%)
1	B maj	B maj	94.51	75.23	+ 3.9	92.63	76.29
2	A maj	A maj	99.45	87.02	+ 13.1	95.48	85.84
3	A min	C maj	99.10	81.62	+ 18.38	95.50	88.42
4	F maj	F maj	99.82	77.22	+ 20.85	95.00	80.99
5	G maj	G maj	79.25	74.13	- 19.79	94.28	83.83
6	D# min	F# maj	70.11	66.97	- 13.79	91.17	77.30
7	F# maj	F# maj	99.22	75.89	+ 23.81	94.73	86.61
8	A maj	A maj	87.84	82.97	- 17.02	94.32	87.96
9	A maj	A maj	82.54	68.37	- 2.74	85.32	71.85
10	C maj	C maj	75.37	70.33	- 3.11	78.30	68.53
			88.72	**75.98**		**91.67**	**80.76**

4.2 Analysis

The per-component analysis of the system that accounts for the errors observed in Tables 1 and 2 is now discussed.

4.2.1 Key Detection

It can be observed that for 2 of the songs (song numbers, 3 and 6 in Table 2), the key has been determined incorrectly. The explanation for this can be based on the theory of the Relative Major/Minor combination of keys[25]. The technique for key determination assumes that the key of the song is constant throughout the length of the song. However, many songs often use both Major and Minor keys, perhaps choosing a Minor key for the verse and a Major key for the chorus, or vice versa. This has a nice effect, as it helps break up the monotony that sometimes results when a song lingers in one key. Often, when switching to a Major key from a Minor key, the songwriters will choose to go to the Relative Major from the Minor key the song is in and vice-versa. This has been taken as a probable explanation for both the songs with erroneous key results where the relative Major has been detected instead of the actual Minor key. Such errors in key recognition do not affect the key filtering as the pitch notes in the Relative Major/ Minor key combination are the same.

4.2.2 Inverse comb filterbank

The inverse comb filters have been used in this implementation for the advantages they seem to offer[17]. Once the filter coefficients are computed, the frequency response of the filter can be easily displayed and checked. The signal filtration can also be done in one pass. Furthermore, the tighter the 'teeth' of the comb are, the more precise the removal can be. However these filters also have some important disadvantages. There is not full control over the design process. The filters exhibit ripples both in passbands and stopbands. Especially the passband ripples (more then 6 dB in some cases) cause distortion during filtration of real musical signals. These signals often exhibit frequency modulation which is converted to amplitude modulation on ripples. In some cases the resulting filter response may be far away from the desired one. Despite having a high order, the filters do not have sufficient stopband attenuation to suppress the harmonics, and the filtration should ideally be done in two or more passes. Finally the design of high order filters with complicated frequency response can also become a very time-consuming process.

4.2.3 Audio Feature

The current implementation uses a simple energy function which calculates the amplitude variation over time in each subband. This is because the vocal frames are normally reflected by a rise in the energy level of the audio. But an analysis solely based on this is often prone to error. For example, a perceptual effect that is predominant in the vocal bands is masking where the high energy of the drums can often partially mask the voice in certain passages. A perceptual evaluation of the residual signal after key filtering highlights a significant attenuation of all the pitched musical instruments except the voice and the drums in the residual signal. We hypothesize that the separation of the voice from other instruments should improve detection accuracy. However from the test results it is observed that the performance improvement obtained by using the filterbank is only marginal. This leads us to infer that the simple energy feature is not optimal to discriminate the voice from other sources of energy.

4.2.4 Text module

The accuracy of the timing information from the text module is dependent on the well-formed nature of the lyrics. That is, being able to decompose every word into its phonemes based on the word's transcription using the CMU Pronouncing Dictionary. The presence of singing without well-formed lyrics, for example, singing with meaningless syllables like 'da', 'uh' will result in the timing error that is observed.

5. DISCUSSION

Based on the per-component analysis discussed above, we are currently investigating various improvements. Comb filters have several disadvantages which have been discussed earlier in this paper. The application of cascaded or parallel connected simple bandstop/bandpass filters has been proven to be a more efficient solution[17]. For the vocal /non-vocal discrimination, more sophisticated features like the spectral contrast proposed in[9] which also consider the spectral peaks, valleys, their difference in each subband and also the relative distribution of the harmonic and non-harmonic components in the spectrum, might serve as a better measure. Text based genre identification[12] and song-specific tempo information could provide valuable information to the text modality. Multiple vocal duration models based on these parameters could be created to enhance the accuracy of duration estimation. Overall, there is considerable room for improvement in the various modules that make up this framework, but the techniques presented in this paper have proven to be capable of musically useful results.

ACKNOWLEDGEMENTS

We thank Dr Min-Yen Kan for testing the vocal duration processor on our current corpus.

REFERENCES

1. Backus, J. *The Acoustical Foundations of Music*, W.W. Norton and Company, December 1977. 2nd edition.
2. Bartsch, M.A. *Automatic singer identification in polyphonic music,* PhD thesis, University of Michigan 2004.
3. Berenzweig, A. and Ellis, D.P.W. "Locating Singing voice segments within music signals ," *Proc. WASPAA* 2001.
4. Berenzweig, A. et al. "Using voice segments to improve artist classification of music," *Proc. AES* 2002.
5. Chou, W. and Gu, L. "Robust singing detection in speech/music discriminator design," *Proc. ICASSP* 2001.
6. Drew, R. *Karaoke Nights: An Ethnographic Rhapsody*, AltaMira Press. November, 2001.
7. Goto, M., and Muraoka, Y. "A beat tracking system for acoustic signals of music," *Proc. ACM Multimedia* 1994.
8. Goto, M. and Muraoka, Y. "Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions," *Speech Communication*, 27(3-4):311–335.
9. Jiang, D.N. et al. "Music type classification by spectral contrast features," *Proc. ICME* 2002.
10. Kato, K. et al. "Blending vocal music with the sound field - the effective duration of autocorrelation function of Western professional singing voices with different vowels and pitches," *Proc. ISMA* 2004.
11. Kim,Y. and Whitman, B. "Singer identification in popular music recordings using voice coding features," *Proc. ISMIR* 2002.
12. Logan, B. et al. "Semantic Analysis of song lyrics," *Proc. ICME* 2004.
13. Maddage, N.C.(a), et al. "Singing voice detection using twice-iterated composite fourier transform," *Proc. ICME* 2004.
14. Maddage, N.C.(b), et al. "Semantic Region Detection in Acoustic Music Signals," *Proc. PCM* 2004.
15. Maddage, N.C.(c), et al. "Content-based music structure analysis with applications to music semantic understanding," *Proc. ACM Multimedia* 2004.
16. Maddage, N.C.(d), et al. "Singer Identification based on vocal and instrumental models," *Proc. ICPR* 2004.
17. Moravec, O. "Comparison of Several Methods for Separation of Harmonic and Noise Components of Musical Instrument Sound," *Proc. International Acoustic Conference* 2002.
18. Morgan, D.P., et al. "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. on Speech and Audio Processing*, September 1997, 5(5): 407-424.
19. Nwe, T.L. and Wang, Y. "Automatic Detection of vocal segments in popular songs," *Proc. ISMIR* 2004.
20. Nwe, T.L. et al. "Singing Voice Detection in Popular Music," *Proc. ACM Multimedia* 2004.
21. Prame, E. "Vibrato extent and intonation in professional Western lyric singing," *JASA*, July 1997, 102(1): 616-621.

22. Rabiner, L.R. and Schafer, R.W. *Digital Processing of Speech signals*, Prentice-Hall, Inc. New Jersey, 1978.

23. Saitou, T. et al. "Extraction of F0 Dynamic characteristics and development of F0 control model in singing voice," *Proc. ICAD* 2002.

24. Seashore, C. E. *Psychology of music*, New York: McGraw-Hill, 1938 & New York: Dover, 1967.

25. Shenoy, A. et al. "Key determination of acoustic musical signals," *Proc. ICME* 2004.

26. Sundberg, J. "The perception of singing," *The Psychology of Music,* San Diego: Academic Press, 1999. 2nd edition. 171-214.

27. Timmers, R. and Desain, P.W.M. "Vibrato: Questions and Answers from Musicians and Science," *Proc. ICMPC* 2000.

28. Tsai, W.H. et al. "Blind clustering of popular music recordings based on singer voice characteristics," *Proc. ISMIR* 2003.

29. Tsai, W.H and Wang, H.M. "Automatic Detection and tracking of target singer in multi-singer music recordings," *Proc. ICASSP* 2004.

30. Tzanetakis, G. "Song-specific bootstrapping of singing voice structure," *Proc. ICME* 2004.

31. Wang, Y. et al. "LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics," *Proc. ACM Multimedia* 2004.

32. Zhang, T. "System and method for automatic singer identification," *Proc. ICME* 2003.