Singer Identification in Popular Music Recordings Using Voice Coding Features

Youngmoo E. Kim
MIT Media Lab
Cambridge, MA 02139
+01 617 253 0619
moo@media.mit.edu

Brian Whitman
MIT Media Lab
Cambridge, MA 02139
+01 617 253 0112
bwhitman@media.mit.edu

ABSTRACT

In most popular music, the vocals sung by the lead singer are the focal point of the song. The unique qualities of a singer's voice make it relatively easy for us to identify a song as belonging to that particular artist. With little training, if one is familiar with a particular singer's voice one can usually recognize that voice in other pieces, even when hearing a song for the first time. The research presented in this paper attempts to automatically establish the identity of a singer using acoustic features extracted from songs in a database of popular music. As a first step, an untrained algorithm for automatically extracting vocal segments from within songs is presented. Once these vocal segments are identified, they are presented to a singer identification system that has been trained on data taken from other songs by the same artists in the database.

1. INTRODUCTION

The singing voice is the oldest musical instrument and one with which almost everyone has a great deal of familiarity. Given the importance and usefulness of vocal communication, it is not surprising that our auditory physiology and perceptual apparatus has evolved to a high level of sensitivity to the human voice. Once we are exposed to the sound of a particular person's speaking voice, it is relatively easy to identify that voice, even with very little training. For the most part the same holds true with regards to the singing voice. Once we become familiar with the sound of a particular singer's voice, we can usually identify the voice, even when hearing a piece for the first time.

Not only is the voice the oldest musical instrument, it is also one of the most complex from an acoustic standpoint. This is primarily due to the rapid acoustic variation involved in the singing process. In order to pronounce different words, a singer must move their jaw, tongue, teeth, etc., changing the shape and thus the acoustic properties of their vocal tract. No other instrument exhibits the amount of physical variation of the human voice. This complexity has affected research in both analysis and synthesis of singing [1].

In spite of this complexity, voice identification is almost effortless to us. But perhaps what is more remarkable is that even in the presence of interfering sounds, such as instruments or background noise, we can still identify the voice of a familiar singer. Thus, our process of identification most likely depends on features invariant to these environmental variations. As will be discussed later, the search for such invariant features that can be used for robust automatic identification is no easy task.

2. BACKGROUND

A significant amount of research has been performed on speaker (talker) identification from digitized speech for applications such as verification of identity. These systems for the most part use features similar to those used in speech recognition. Many of these systems are trained on pristine data (without background noise) and performance tends to degrade in noisy environments. And since they are trained on spoken data, they perform poorly to singing voice input. For more on talker identification systems, see [2].

In the realm of music information retrieval, there is a burgeoning amount of interest and work on automatic song and artist identification from acoustic data. Such systems would obviously be useful for anyone attempting to ascertain the title or performing artist of a new piece of music and could also aid preference-based searches for music. Another area where this research has generated a great deal of interest is copyright protection and enforcement. Most of these systems utilize frequency domain features extracted from recordings, which are then used to train a classifier built using one of many machine learning techniques. Robust song identification from acoustic parameters has proven to be very successful (with accuracy greater than 99% in some cases) in identifying songs included in the database [3]. Artist identification is a much more difficult task, and not as well-defined as individual song identification. A recent example of an artist identification system is [4], which reports accuracies of approximately 50% in artist identification on a database of about 250 songs.

Also relevant to the task of singer identification is work in musical instrument identification. Our ability to distinguish different voices (even when singing or speaking the same phrase) is akin to our ability to distinguish different instruments (even when playing the same notes). Thus, it is likely that many of the features used in automatic instrument identification systems will be useful for singer identification as well. Work by Martin [5] on solo instrument identification demonstrates the importance of both spectral and temporal features and highlights the difficulty in building machine listening systems that generalize beyond a limited set of training conditions.

Obviously, singer identification and artist identification can amount to the same thing in many situations. In [6], Berenzweig and Ellis use vocal music as an input to a speech recognition system, achieving a success rate of 80% in isolating vocal regions. In [7], Berenzweig, Ellis, and Lawrence use a neural network trained on radio recordings to similarly segment songs into vocal and non-vocal regions. By focusing on voice regions alone, they were able to improve artist identification by 15%.

The system presented here also attempts to perform segmentation of vocal regions prior to singer identification. After segmentation, the classifier uses features drawn from voice coding based on Linear Predictive Coding (LPC), although with some modification. LPC is particularly good at

highlighting formant locations (regions of resonance, which have been shown to be especially significant perceptually [8]). Much research has been performed on speech coding via LPC, and its uses are ubiquitous today. For example, all digital cellular phones use LPC-based voice coders and maintain fairly good sound quality, even at low information rates. Since these models were designed to primarily carry the human voice accurately, it seems logical that they could be useful for singer identification as well.

3. DETECTION OF VOCAL REGIONS

Before trying to establish who is singing a particular piece, it is obviously important to identify the sections of that piece in which singing actually occurs. In this section, we present a technique for automatically detecting these regions of singing within recordings.

3.1 Vocal frequency regions

The majority of energy in the singing voice falls between 200 Hz and 2000 Hz (with some variation depending on the singer). This is a region that the human auditory system is particularly sensitive to. Classic experiments on a wide variety of human listeners have established general equal-loudness curves [9] which establish that signals within this mid-range of frequencies are perceived as louder than signals of equivalent absolute amplitude at other frequencies, higher or lower.

Another perceptual effect that predominates in this region is masking, in which energy in one frequency band will obscure or "mask" lesser energies in adjacent frequency bands. In most recorded vocal music, the tendency is to isolate other instruments away from the voice so as not to mask or be masked by the voice. One notable exception is singing with a symphony orchestra. Many orchestral instruments fall in the same frequency range as the voice, and the sheer number of instruments is more than enough to mask the untrained voice. Classically trained singers, however, are taught to produce extra resonance at a higher frequency (often referred to as the *singer's formant*, located around 2500 Hz), which allows the voice to be perceived against the overwhelming numbers [10]. But since the majority of popular recorded music is not in this style, we can restrict our range of interest to lower frequencies.

Since we are interested in detecting regions of singing, a straightforward method would be to detect energy within the frequencies bounded by the range of vocal energy. A very simple approach is to filter the audio signal with a band-pass filter which allows the vocal range to pass through while attenuating other frequency regions. For this, we use a simple Chebychev infinite-impulse response (IIR) digital filter of order 12. The frequency response of this filter is shown in Figure 1.

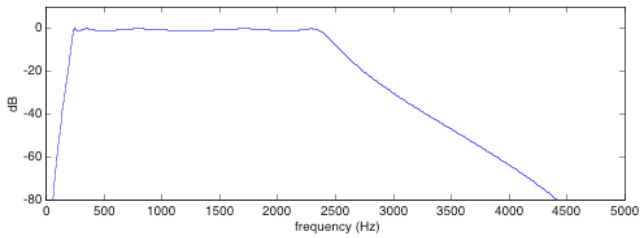

Figure 1: Vocal enhancement filter frequency response.

This filter has the musical effect of attenuating other instruments that fall outside of this frequency region, such as bass and cymbals. But even in popular music, the voice is not the only instrument producing energy in this region. Drums, for example, disperse energy over a wide range of frequencies, a significant amount of which falls in our range of interest. So another measure is needed to discriminate the voice from these other sources.

3.2 Detection via harmonicity

Singing primarily consists of sounds generated by phonation, the rapid vibration of the vocal folds resulting in utterances referred to as *voiced* by speech researchers. 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]. Singing is >90% voiced, whereas speech is only ~60% voiced [11]. Because of this, the singing voice is highly harmonic (energy exists at integer multiples of the fundamental frequency, or pitch). Other high energy sounds in this region, drums in particular, are not as harmonic and distribute their energy more widely in frequency.

To exploit this difference, we use an inverse comb filterbank to detect high amounts of harmonic energy. The block diagram and frequency response of a simple inverse comb filter is shown in Figure 2. By passing the previously filtered signal (as described above) through a bank of inverse comb filters with varying delays, we can find the fundamental frequency which the signal is most attenuated. By taking the ratio of the total signal energy to the maximally harmonically attenuated signal, we have a measure of *harmonicity*, or how harmonic the signal is within the analysis frame.

$$H = \frac{E_{original}}{\min_i \left(E_{filtered,i} \right)} \quad (1)$$

By thresholding the harmonicity against a fixed value, we have a detector for harmonic sounds. The hypothesis is that most of these correspond to regions of singing. Results using this technique are presented in Section 5.

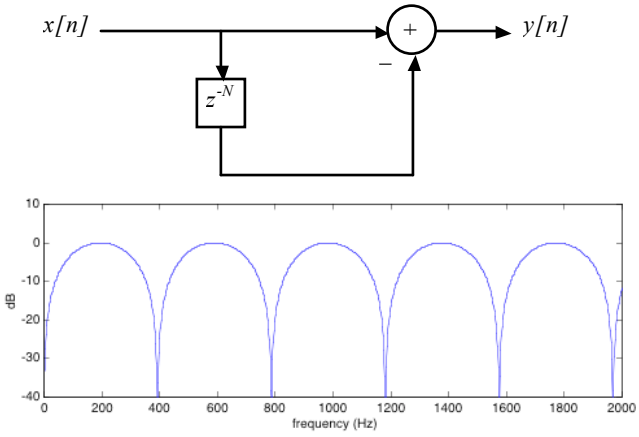

Figure 2: Block diagram of the simple inverse comb filter (top). The frequency response of an inverse comb filter, tuned to ~400 Hz (bottom). The spacing of the attenuated frequencies is determined by the delay parameter *N*.

4. SINGER IDENTIFICATION

Although relatively easy for humans, robust singer identification is an extremely difficult task for a machine listening system. Even with clean signals (with no other instruments or background noise), simple frequency or time domain features do not lead easily to a unique "voiceprint". And in most performances or recordings where the voice is amidst a mixture of other sounds, the problem becomes even more complex. This section discusses the features, extraction methods, and classification techniques used in the singer ID system.

4.1 Features from Speech Coding

Much research (primarily dealing with speech) has focused on voice coding using an analysis/synthesis approach. In this approach a source signal is analyzed and re-synthesized according to a source-filter model of the human voice. This is the general principle behind Linear Predictive Coding (LPC). The primary advantage of this technique has been its utility in compressing speech data resulting in the low-bitrate speech coders used in many applications today.

4.1.1 Traditional LPC

The goal of linear predictive analysis is to establish an estimate, $\tilde{s}[n]$ to the source signal $s[n]$, using a linear combination of p past samples of the input signal:

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{2}$$

The coefficient values α_k in Equation (2) are determined by minimizing the mean squared prediction error, which is the difference between the source signal and the predicted signal:

$$E_m = \sum_m \left(s_m[n] - \tilde{s}_m[n]\right)^2 \tag{3}$$

The transfer function relating the source signal and the signal estimate is shown [12] to be an all-pole filter:

$$H[z] = \frac{G}{A[z]} \tag{4}$$

where the denominator is defined as follows:

$$A[z] = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{5}$$

This demonstrates how linear predictive analysis is equivalent to a source-filter model, where the vocal tract response is modeled using a time-varying all-pole filter function of order p. The calculated coefficients can be factored to determine the pole locations, which generally correspond to the formants (resonances) of the vocal tract. An example of an LPC filter response is shown in Figure 3.

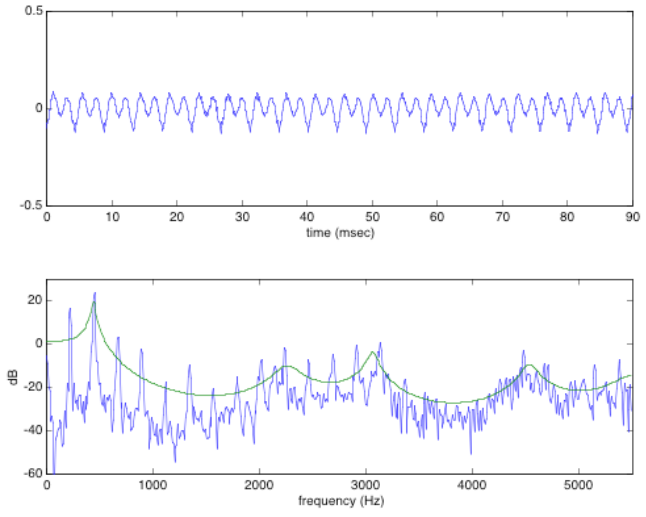

Figure 3: A vocal segment (top) and its spectrum (bottom). The line in the bottom figure shows the 12-pole LPC estimate. The peaks indicate the formant locations.

As regions of high energy within a frequency range of enhanced sensitivity, vocal formants are of special perceptual significance. The general pattern of formant locations determines our perception of phonemes, and thus language. It is also believed that an individual's particular formant patterns are a key feature for speaker identification [8]. Due to increased harmonic energy in singing, formants are enhanced, so it is reasonable to believe that they would be important for singer identification as well. Formant frequencies and magnitudes extracted via LPC (with 12 poles) are used as key features in the singer classifier.

A common technique for minimizing the prediction error (Equation 3) uses the autocorrelation matrix (hence its name, the autocorrelation method), which can be calculated from the power spectrum of the signal. This will be useful in the next section, which discusses warping of the power spectrum to better fit established theory on auditory perception.

4.1.2 Warped Linear Prediction

One disadvantage of standard LPC is that it treats all frequencies equally on a linear scale. However, the human ear is not equally sensitive to all frequencies linearly. In fact our frequency sensitivity is very close to logarithmic. As a result, LPC systems sometimes place poles at higher frequencies where the ear is less sensitive and miss closely spaced resonant peaks at lower frequencies where the ear is more sensitive. Using a higher order LPC is one way of compensating for this, but increasing the number of poles makes it difficult to track correlations between analysis frames.

Instead, we use a *Warped Linear Prediction* model by pre-warping the power spectrum of each frame [13], [14]. The warping function can be made to closely approximate the Bark scale, which approximates the frequency sensitivity of human hearing. Warping is achieved through the following relation:

$$\hat{\omega} = \omega + 2\tan^{-1}\left(\frac{a\sin\omega}{1 - a\cos\omega}\right) \tag{6}$$

A parameter value of a=0.47 closely matches the Bark scale. This allows us to keep the order of the analysis low (and thus more easily track formants from frame to frame) while more accurately capturing formant locations, especially at lower frequencies. The additional emphasis on lower frequencies has the added benefit of being able to pick out individual low harmonics. Standard LPC does not have the resolution at low frequencies to detect individual harmonics, but warped LPC is oftentimes able to identify the lowest harmonic, which usually corresponds to the pitch, which can be a useful feature in singer identification. Figure 4 shows an example of warped LP.

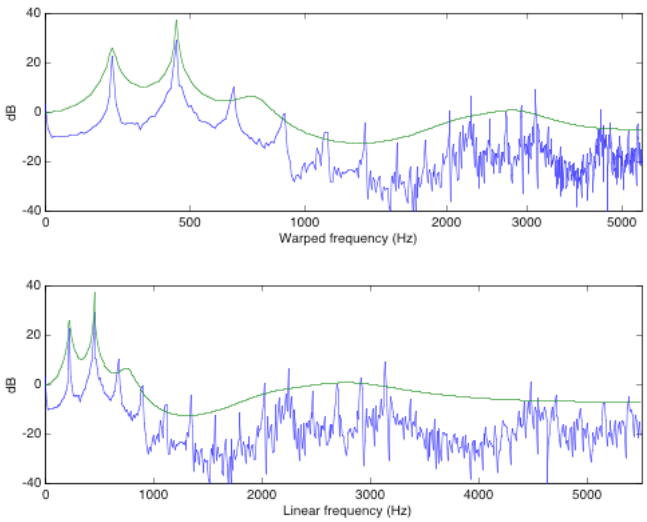

Figure 4: The same vocal segment analyzed via warped Linear Prediction. The warped frequency scale is on top and the unwarped scale on the bottom. Note the particularly good fit at lower frequencies.

4.2 Classification Techniques

Two different classifiers were trained using established pattern recognition algorithms. A brief description of the two classifiers implemented follows. In this task, each "class" represents an individual singer.

4.2.1 Gaussian Mixture Model (GMM)

The Gaussian mixture model uses multiple weighted Gaussians to attempt to capture the behavior of each class of training data. The use of multiple Gaussians is particularly beneficial when analyzing data that has a distribution not well modeled by a single cluster. It is a very flexible model that can adapt to encompass almost any distribution of data. Test points are classified by a maximum likelihood discriminant function, calculated by their distances from the multiple Gaussians of the class distributions [15].

To determine the parameters of the Gaussians that best model each class, we use the well-known technique of *Expectation Maximization* (EM). EM is an iterative algorithm that converges on parameters that are locally optimal according to the log-likelihood function. Thus, it is sensitive to initial conditions, and usually several runs are performed to ensure that the derived fit is a relatively good one. It is also useful to perform *Principle Components Analysis* (PCA) prior to EM. PCA is a multi-dimensional rotation of the data onto the axes of maximal variance. It also has the added benefit of normalizing the data variances, which avoids highly different scaling among the dimensions, which is problematic for EM.

The number of Gaussians used is generally a user-defined parameter, and some experimentation is usually required to find a reasonable number for a given data set.

4.2.2 Support Vector Machine

A Support Vector Machine (SVM) is based on statistical error minimization techniques applied to a machine learning domain [16]. SVMs work by computing an optimal hyperplane that can linearly separate (in one case) two classes of data. These hyperplanes simplify to a set of Lagrange multipliers for each training case, and the set of points within the dimensional vectors fed for training that have non-zero Lagrangians are the *support vectors*. The machine saves these support vectors and applies them to new data in the form of the test set for further on-line classification.

We used an SVM with a Gaussian kernel. Our C (maximum lagrangian value) was set to 10. Each class was trained as a separate SVM with all the positive examples from that class along with the same amount of randomly chosen negative examples. After training each individual SVM, we applied the confidence thresholding metric discussed in [4] to remove uncertain frame classifications and chose the SVM with the highest confidence to classify frames in the test set [17].

5. EXPERIMENT AND RESULTS

Several different experiments were conducted using the features and algorithms described in sections above. First, the data set used for training and testing is described. Then the specifics regarding the experimental configuration and methods are presented, followed by the reporting of experimental results.

5.1 The Data Set

The data sets used in this experiment are various subsets of the NECI Minnowmatch testbed [4] with some minor additions. The entire Minnowmatch testbed consists of >250 songs from the albums of more than 20 distinct artists/groups. Some of these songs, however, do not contain singing, and were eliminated from consideration. Additionally, on occasion different individuals sing some of the songs performed by a single group. Care was taken to classify each song by the actual singer on the recording. After these considerations, the resulting testbed included 17 different solo singers and slightly more than 200 songs. All songs were downsampled to 11.025 kHz (from the CD sampling rate of 44.1 kHz) to reduce the data storage and processing requirements. Even at the lower sampling rate, most vocal energy falls well below the Nyquist rate (half the sampling-rate) and is preserved for our analysis.

5.2 Experimental Procedure and Results

5.2.1 Detection of Vocal Regions

To test the accuracy of the vocal segment detector, a subset (20 songs, or approximately 10%) of the database was segmented manually "by ear" into regions of singing and non-singing to provide a set of "ground truth" data. The relatively small size of the ground truth data set is because of the tedious and somewhat ill-defined nature of this task. Establishing exactly where a vocal segment begins and ends with certainty is problematic. Low-level background vocals that tend to fade in and out in some songs add further complications. The segmentation on this set was as accurate as can be, given the difficulties.

This set of 20 songs was then analyzed using the vocal detection system described in Section 3. Analysis was performed using frame size of 1024 samples (~100 msec) and frames were taken every 512 samples (~50 msec). The initial threshold for vocal classification was a harmoncity value of H>2. As can be seen from the results in Table 1, the classifier is not very accurate, though it does perform better than chance (two classes = 50%). As a front-end for the singer identification system, however, we would like to avoid false positives (identifying regions of non-vocal music as having vocals), since the identity classifier would attempt to place these non-vocal segments with a singer anyway. On the other hand, false negatives (classifying regions of vocals as being instrumental only) are more acceptable since they simply reduce the amount of data fed to the singer ID system. Given this tradeoff, we can greatly reduce the number of false positives by raising the threshold of H, at the expense of more false negatives. As shown in Table 3, raising the threshold to H=2.6 reduced the error rate among non-vocal frames to ~20% while retaining ~30% of the vocal frames. This value was used to automatically segment the data in the singer identification system.

Table 1: Performance of vocal detector at multiple harmonicity thresholds.

H Threshold	Vocal Segments	Non-vocal Segments	All Segments
2.0	55.4%	53.1%	55.4%
2.3	40.5%	69.2%	55.1%
2.6	30.7%	79.3%	54.9%

5.2.2 Singer Identification

For singer classification, we used approximately half of the database (the odd numbered songs from albums) to train the classifier and the remaining songs to evaluate the performance of our classifier. We conducted two sets of experiments: In the first set, LPC features were extracted from entire songs and used for classification. The second set of experiments used only features from regions classified as containing vocals. In both experiments, analysis frames were again 1024 samples calculated at 512 sample intervals. A 12-pole LP analysis was performed on both linear and warped scales. The frequencies and magnitudes of the pole locations were used as inputs to the classifiers.

Three different feature sets (linear scale data, warped scale data, and both linear and warped data) were tested, and two different classifiers (GMM and SVM) were used in each case. In the SVM classifier, only every tenth data frame was used because of computer memory constraints. The results from these experiments are summarized in Table 2. The highest performing number of Gaussians was used as the GMM result.

Table 2: Singer classification results. Results are listed as percentages of songs correctly classified (followed by percentages of individual frames correctly classified).

Experiment 1: Entire song data

Features	GMM	SVM
Linear frequency features	32.1 (16.6)	39.6 (30.7)
Warped frequency features	31.3 (17.1)	35.0 (30.4)
Linear and warped features	33.4 (16.5)	45.3 (29.6)

Experiment 2: Only song segments classified as vocals

Features	GMM	SVM
Linear frequency features	36.7 (18.1)	35.8 (17.6)
Warped frequency features	33.0 (17.4)	34.0 (26.8)
Linear and warped features	38.5 (16.6)	41.5 (28.8)

On the whole, the classification results are far greater than chance (17 classes = ~6%), but still fall well short of expected human performance. In general, the linear frequency features tend to outperform the warped frequency features when each is used alone, but using them together does benefit performance. Strangely, song and frame accuracy increases when using only vocal segments in the GMM, but decreases in the SVM using the same segments.

6. DISCUSSION AND FUTURE WORK

As shown in the results, the raw accuracy of the vocal detector could use some improvement. But given the uncertainties in the defining the exact boundaries of vocal and instrumental segments, the raw output numbers have some uncertainty attached as well. It should be noted that the singing detector presented here is an untrained system (it possesses no prior knowledge). Other systems ([6] and [7]) have achieved higher vocal detection accuracy (as well as [18] for speech vs. music), but have been trained on ground-truth databases. It would be possible to combine features from both systems to achieve greater accuracy. A better perceptual model (than the static filter used here) may be of substantial benefit as well. We are currently investigating these possible improvements.

Qualitative listening to the detected regions demonstrates some interesting points. Many of the extended mislabeled regions occur at the beginning of songs or during instrumental bridges where producers have highlighted instruments in the absence of vocals. Also, the beginnings and ends of phrases tend to be cut off, since they contain less harmonic content. Regions of extended vowels (such as held notes) are particularly well detected. A simple extension to the system would be to pad each section with extra time on either side. But whether this would aid in singer identification is an entirely different question. It is conceivable that higher-level musical knowledge could be added to the system in an attempt to identify song structure, such as the location of verses and choruses, from patterns in the segmentation data. The probability of vocals in those sections could be weighted more strongly than in others, which would reduce the problem of falsely classifying strong solo instrumental passages as vocals.

The singer classification results are notable in a few ways. Both classification techniques found something discriminatory within the features. With the GMM, accuracy increased as expected when the classifier was trained and tested with the voice-classified frames. That the performance of the SVM decreased is a bit puzzling. It is likely that the SVM is finding aspects of the features that are not specifically related to the voice. The fewer number of frames in the second experiment might account for the decrease in accuracy. There is also a great deal of uncertainty about the accuracy of the voice labeling, and that is likely factor as well.

The better performance of the linear frequency scale features vs. the warped frequency features probably indicates that the machine finds the increased accuracy of the linear scale at higher frequencies useful. Though this is contrary to human auditory perception, it is not surprising that there is discriminatory information there, though it is probably not correlated with the voice. The increased performance in using both linear and warped features indicates that the analyses are not completely redundant.

Given the relatively low frame accuracy reported in the singer identification experiments, the overall frame confusion matrix (Figure 5) is not surprising. There is not a high amount of intensity along the diagonal, except for a few particular artists. It is as yet unclear why these particular singers are easily detected. There may have been particular qualities to these voices or other parts of the songs, which may have been highlighted by the linear predictive analysis. This remains an ongoing investigation in our research.

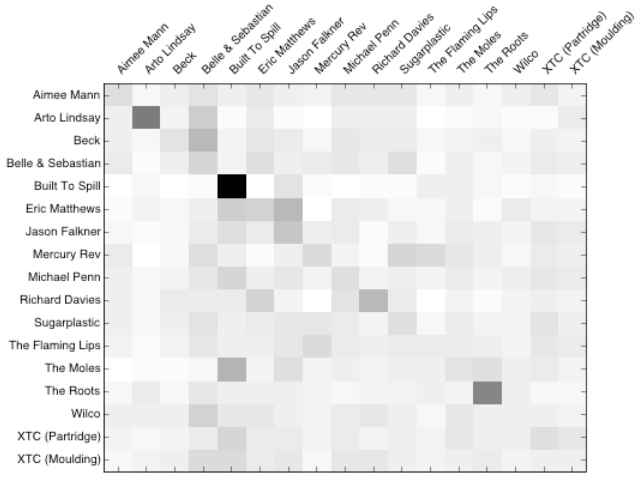

Figure 5: Confusion matrix for all voiced data frames, using both linear and warped features.

Others (e.g. [4] and [7]) have achieved higher artist classification accuracy on the same test data, but have not attempted to identify the individual singers. In [7], using only vocal segments improved accuracy in a neural net classifier using Mel-frequency Cepstral Coefficients (MFCCs). But it is unclear in that study (and in the research presented here as well) whether the classifier is actually training on vocal features or is using some other aspect of the recordings, even though both MFCCs and linear prediction have proven useful in speech applications. Their system also uses the differences between MFCCs as features, indicating that using differential magnitudes (in addition to or instead of raw magnitudes) may be beneficial.

Another possibility for improving the singer classifier is to incorporate more time-varying information. While almost all audio analysis is conducted at a fixed rate, the actual information rate of audio varies widely. For this reason, a state-based classifier, such as a Hidden Markov Model, may improve

classifier performance. This will be explored in our research at a later date.

7. ACKNOWLEDGEMENTS

Our thanks to Dr. Paris Smaragdis for his help in enlightening us to some of the more obscure aspects of pattern recognition. And thanks to Prof. Barry Vercoe for his support of our research.

8. REFERENCES

[1] Y. E. Kim, "Excitation Codebook Design for Coding of the Singing Voice," *Submitted to the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001.

[2] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 58-71, 1996.

[3] J. Herre, E. Allamanche, and O. Hullmuth, "Robust Matching of Aduio Signals Using Spectreal Flatness Features," presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2001.

[4] B. Whitman, G. Flake, and S. Lawrence, "Artist detection in music with Minnowmatch," presented at *IEEE Workshop on Neural Networks for Signal Processing*, Falmouth, MA, 2001.

[5] K. Martin, *Sound-Source Recognition: A Theory and Computational Model*. Ph.D. Thesis. Massachusetts Institute of Technology, Cambridge, MA, 1999.

[6] A. L. Berenzweig and D. P. W. Ellis, "Locating Singing Voice Segments Within Music Signals," presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2001.

[7] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using Voice Segments to Improve Artist Classification of Music," In press, 2002.

[8] R. Brown, "An experimental study of the relative importance of acoustic parameters for auditory speaker recognition," *Language and Speech*, vol. 24, pp. 295-310, 1981.

[9] H. Fletcher, "Auditory patterns," *Review of Modern Physics*, vol. 12, pp. 47-65, 1940.

[10] J. Sundberg, *The Science of the Singing Voice*. Dekalb, IL: Northern Illinois University Press, 1987.

[11] P. R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. Ph.D. Thesis. Stanford University, Stanford, CA, 1990.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[13] H. W. Strube, "Linear prediction on a warped frequency scale," *Journal of the Acoustical Society of America*, vol. 68, pp. 1071-1076, 1980.

[14] A. Härmä, "A Comparison of Warped and Conventional Linear Predictive Coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 579-588, 2001.

[15] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2000.

[16] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 955-974, 1998.

[17] G. Flake, "NODElib,".: NEC Research Institute.

[18] E. D. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," presented at *IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, 1997.