USING VOICE SEGMENTS TO IMPROVE ARTIST CLASSIFICATION OF MUSIC

ADAM BERENZWEIG[1], **DANIEL P. W. ELLIS**[2], **AND STEVE LAWRENCE**[3]

[1,2]*Department of Electrical Engineering, Columbia Univerity, New York, NY 10027*
alb63@columbia.edu
dpwe@ee.columbia.edu
[3]*NEC Research Institute, Princeton, NJ 08540*
lawrence@necmail.com

Is it easier to identify musicians by listening to their voices or their music? We show that for a small set of pop and rock songs, automatically-located singing segments form a more reliable basis for classification than using the entire track, suggesting that the singer's voice is more stable across different performances, compositions, and transformations due to audio engineering techniques than the instrumental background. The accuracy of a system trained to distinguish among a set of 21 artists improves by about 15% (relative to the baseline) when based on segments containing a strong vocal component, whereas the system suffers by about 35% (relative) when music-only segments are used. In another experiment on a smaller set, however, performance drops by about 35% (relative) when the training and test sets are selected from different albums, suggesting that the system is learning album-specific properties possibly related to audio production techniques, musical stylistic elements, or instrumentation, even when attention is directed toward the supposedly more stable vocal regions.

INTRODUCTION

A current research topic of interest to many who apply machine learning techniques to audio and music is the search for musically relevant features. The methods of statistical pattern recognition can only realize their full power when real-world data are first distilled to their most relevant and essential form. In the case of speech recognition, low-level spectral features such as cepstral coefficients have proven sufficient. However, for applications that require a more sophisticated high-level picture of the data, the question remains open. For example, when building a music recommendation engine, we might ask, "What is it about the music you like that makes you like it? And, what sort of features might capture that essential quality?" In an audio-based music retrieval system, we look for features that are invariant to the kinds of acoustic-level transformations applied during audio production, implying that simple spectral features will not suffice.

In this paper, we consider the question of whether certain parts of songs are more discriminatory than others. Specifically, we look at vocal segments versus instrumental music segments, and ask whether the singer's voice is more distinctive than the instrumental background as a basis for artist classification. The question has relevance to recommendation engines (perhaps it is the quality of certain artists' voices that you like), content-based music information retrieval (IR), and the general question of how to capture the essential character of a piece of music with a few hundred numbers.

We use an artist classification task to explore these issues. It must be noted that this task is inherently ambiguous, since the "artist" responsible for a song could refer to the composer, the performer, the producer, the audio engineer, the singer, the guitarist, and so on. Many people are involved with the creation of a professional recording, and they all leave their mark. However, for rock and pop, the composer and performer is often the same, and the performer is typically more recognizable than the composer. For our purposes the band or performing musician is treated as the artist.

1. APPROACH

Our basic paradigm is to classify musical tracks as being the work of one of several predefined artists. In all cases, we calculate features from the audio content, use a statistical classifier to estimate the artist based on short temporal contexts of those features, then combine the estimates from the entire track to generate the final artist classification.

The experiments are designed to show the effect of using voice- or music-only segments on an artist classification task. First we obtain a baseline result using the unsegmented audio. Then we attempt to locate segments of the music that are dominated by singing voice, using a straightforward classifier trained for that task. Note that the voice is only segmented in time, not separated in frequency, and thus the segments returned will at best still be mixtures of voice and music, ideally with the voice prominent in the mixture. By the same token, the segments not selected by this first stage of classification should consist of instruments without any vocals, to the extent that the singing-segment classifier is successful. Once the voice segments have been identified, the artist classifier is run again using only the extracted segments. For comparison, we also perform classification using only the segments identified as non-voice—roughly speaking, the complement of the previous case, although certain frames may appear in neither set due to smoothing and a minimum-duration constraint, or when using dynamic thresholding.

There are several options for how to use the segments to aid the classifer. In this paper we describe two methods, "posterior segmentation" and "retraining". In the posterior segmentation method, the artist classifier is trained on unsegmented data, but the final artist classification is based only on estimates relating to frames identified as voice (or music) by the segmentation system. In the retraining method, both the training and test sets are segmented, and the artist classifier is retrained on the segmented data (e.g., vocals only). In testing, only segments of the matching vocals/music class are used, all of which contribute to the final artist classification.

The details of the experimental setup are provided in the next section.

2. EXPERIMENTS

2.1. Vocal Segmentation

To segment voice from instrumental music, a two-class (voice/music) multi-layer perceptron (MLP) neural net was trained using hand-labelled data. The features used are 13 perceptual linear prediction (PLP) coefficients, calculated over a 32ms window hopped every 16ms. Deltas and double deltas are added for each frame, and 5 consecutive frames are presented to the network simultaneously to provide context. Thus the total number of input units to the network is $13 * 2 * 5 = 130$. A single hidden layer with 50 units was used. The output layer had two units, for voice and music.

The neural network was trained using the QuickNet software from ICSI Berkeley, using back-propagation with a minimum-cross-entropy criterion. A simple, fixed learning-rate decay scheme with online updates over 9 epochs was adopted based on initial experiments to avoid overfitting.

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 (voice and music). The stream of interest is smoothed with a simple moving-average filter of length 40 frames (640ms), then compared to a threshold. Frames whose probability exceeds the threshold are selected for segmentation. Finally, a minimum duration constraint is enforced by discarding segments whose length is smaller than a certain amount. For the results reported below, a minimum duration of 10 frames (160ms) was chosen. The value of the threshold, and hence the number and expected "purity" of the identified frames, was varied depending on the experiment, as described below.

The data used to train the segmentation network is distinct from the data used in later experiments, and in fact has slightly different characteristics which may have hurt the performance of the segmentation system. This system was trained on a series of (mono, 22.05kHz) 15-second clips recorded from the radio by the authors of [1], whereas the training and test data for the artist classification task described in this paper are taken from CDs (and then downsampled and mixed to mono, 22.05kHz). We chose to train the segmenter on the Scheirer-Slaney data because we had previously hand-labelled it as voice/music for earlier work.

In [2], Berenzweig and Ellis describe a more sophisticated approach to segmenting vocals from music using a speech recognizer as a "voice detector". However, we used the simpler approach above due to computational reasons. This approach is sufficient for the purposes of this paper, where we focus on comparing the use of voice and music segments for classification. We do not need to identify all vocal segments, we only need to identify a sufficient percentage with a low error rate. This can be accomplished by finding a point on the ROC curve of the segmenter system where the false alarm rate is sufficiently low. The ROC curve of our vocals detector, evaluated on data representative of our artist classification task, is shown in Figure 1. By choosing a threshold of 0.55, we can obtain about 40% of the true vocal frames, while keeping the false alarm rate at around 10%.

In other experiments, instead of choosing a fixed threshold based on the ROC curve, we use a *dynamic* threshold that adjusts to select a fixed percentage of frames from each track, in this case 25%. The segmentation routine searches for a threshold that selects about 25% of the frames, after application of the minimum duration constraint. The rationale for this alternative is that some of the music may contain little or no vocals (or few frames that our vocals classifier can recognize), defeating the artist classification approach using vocal segments. In this situation, our solution is to use the most voice-like frames for classification, even if they are not in fact voice. In the case where more than 25% of a track would have been selected as voice by the fixed threshold, the dynamic threshold ensures that we use only the segments most confidently identified as voice as the basis for classification.

2.2. Artist Classification

The artist classification is also performed by an MLP neural network. The input is again a vector of cepstral coefficients, but in this case we use 20 mel-frequency cepstral

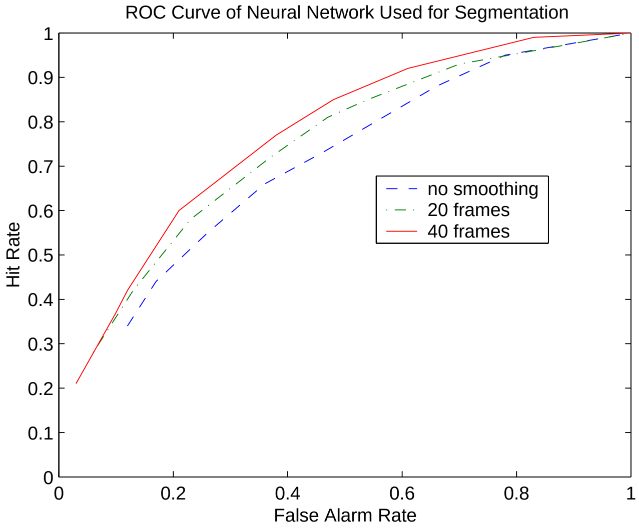

Figure 1: ROC curve for the neural network trained to segment singing voice from instrumental music. For the fixed threshold segmenter, the operating point was chosen to keep the false alarm rate at 10% while obtaining about 40% of the true vocal frames.

coefficients (MFCCs) instead of the PLP cepstra used for segmentation. In our previous work on speaker identification, we found MFCCs were much better for distinguishing between speakers [3]; PLP coefficients are oriented towards speech recognition—recognizing the words regardless of the speaker—and were specifically developed to suppress the differences between individual speakers, which they do quite effectively. Clearly, for the artist classifier, this is not the behavior we desire. MFCCs have been used with success in other speaker identification tasks [4] because they capture more of the characteristic qualities of individual voices. (We have yet to evaluate whether our vocals detection classifier would perform better if based on MFCCs, but that task is less dependent on small differences between voices). Recent research suggests that MFCCs are appropriate for modeling music as well as speech [5].

The 20 MFCCs and 20 first-order deltas are combined and presented to the network, again using a context window of 5 frames. Thus the network input size is 200, and we use a hidden layer of 50 units. The network is trained according to the same procedure used for the vocals detection net. The size of the output layer varies between five and 21, depending on number of artists in the data set as described in the next section.

2.3. Retraining vs. Posterior Segmentation

In section 1, we described two approaches to using segments identified as particularly discriminant (e.g., the vocals) to aid in overall classification. The straightforward scheme of posterior segmentation trains the artist classifier on entire tracks, but then ignores everything but the discriminant segments when calculating the overall classification of a track. This was our initial approach, and performed reasonably well. We were able to obtain significant improvements in accuracy, however, by instead adopting a full retraining scheme: Once a detection scheme is defined for locating the vocal (or instrumental) segments, the artist classifier is retrained on those segments alone. This results in a classifier more specialized for classification of the segments in question, and thus the observed performance improvements of around 10% relative are in line with expectations. The disadvantage of this approach, when compared to posterior segmentation, is that new artist classifier networks must be trained for each possible segmentation scheme—in this case, the whole track, the segments tagged as vocals, and the non-vocals segments.

2.4. Data sets

We use several different data sets representing different mixtures of artists. The largest set contains 269 full songs from 21 different artists (one album of 10-15 songs per artist), comprising 1057 minutes of audio. From each album, approximately half of the songs are allocated to the training set, and the remainder to the test set; the songs in each set are entirely disjoint. 135 songs are used for training, and 134 for testing.

The 21 artists represent a variety of styles of rock and pop music. We use the same data set as Whitman et al. in [6], and the artists are listed in Table 1. Five of the artists are electronic musicians that do not use vocals, making them particularly difficult for our approach. To obtain a better indication of the approach's capability in the domain for which it is intended, a second data set of 16 artists is defined excluding the nonvocal electronic musicians. Finally, a data set of five artists for which we had at least two albums was selected to highlight the "album effect", as described in the next section.

2.5. The "Album Effect"

Especially in popular music, it is very common for all the recordings on a single album to share many common elements, such as audio production techniques, stylistic themes, and instrumentation. It is conceivable that the classifier might learn to recognize *albums* and not artists. For example, the overall spectral shape of the record is formed by the effects processing, equalization, and compression that the engineer chooses to apply when mixing and mastering the album. It is plausible that this signature spectral shape is the easiest way to distinguish between artists, particularly if the basis for classification is cepstral coefficients over a small window. Similarly, in terms of instrumentation, use of strings versus horns may

	Data Set Size		
Artist	21	16	5
Aimee Mann	X	X	
Arto Lindsay	X	X	
Beck	X	X	
Belle & Sebastian	X	X	
Boards of Canada	X		
Built to Spill	X	X	
Cornelius	X		
DJ Shadow	X		
Eric Matthews	X	X	
Jason Falkner	X	X	
Mercury Rev	X	X	
Michael Penn	X	X	X
Mouse on Mars	X		
Oval	X		
Richard Davies	X	X	
Roxette			X
Stereolab			X
Sugarplastic	X	X	
The Flaming Lips	X	X	
The Moles	X	X	
The Roots	X	X	
Wilco	X	X	X
XTC	X	X	X

Table 1: Artists and data sets.

leave a spectral signature on an album.

To examine the album effect and its role in artist classification, we formed two additional data sets of 5 artists. Two albums from each of the 5 artists are used. In the first set, which we shall call **5-matched**, both the training and the test sets contain songs from both albums. In the second set, **5-mismatched**, the classifier is trained on one album from each artist, then tested on the other. In this way, we can distinguish between the influence of album-specific cues, and the more difficult case of recognizing artists even when the cues are different.

2.6. Evaluation

We use two metrics to evaluate the system: frame accuracy and song accuracy. Frame accuracy measures the percentage of input frames which are labeled by the classifier with the appropriate artist class. The output from the classifier is then combined by a summarization routine which makes a single guess per song at the identity of the artist based on the frame-level posterior probabilities of the artist classes over the duration of the entire song. The accuracy of this labeling is the song accuracy.

The summarization routine calculates a confidence-weighted average of the posterior probabilities of each artist class over all of the selected segments (all, vocals or nonvocals) for each track. Several different confidence measures were tested, such as inverse entropy of the frame (treating the vector of posterior probabilities as a probability distribution), the margin (difference between the winning class posterior and the 2^{nd} highest posterior), and the value of the largest posterior. Additionally, we implemented a minimum confidence threshold, such that frames whose confidence did not exceed the threshold were ignored. Averaging of the confidence-weighted posterior values was performed in an exponentiated domain, where the exponent L was varied between 1 (for a simple mean) to 10 as an approximation to the infinity-norm, which is dominated by the largest single posterior across all frames being averaged. Unless noted otherwise, the summarization method used for the experiments reported below is simply the mean posterior probability over the entire song (i.e., no confidence weighting).

3. RESULTS AND DISCUSSION

The results are summarized in Table 2. Overall, these results compare favorably to the previously-reported result of Whitman et al. [6], who presented an artist classification system based on neural networks and support vector machines. On the same 21-artist dataset, they reported a peak accuracy of 50% at the song level compared to our best result of 64.9% (Figure 2).

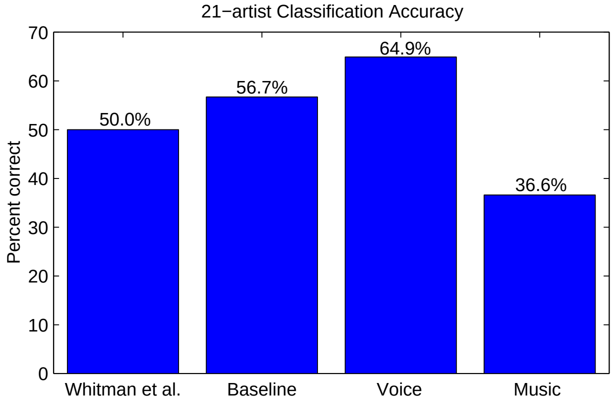

Figure 2: Summary of results for the largest dataset (21 artists, 269 songs). Our baseline result using the entire songs improves on the previous result using this dataset by Whitman et al. [6] (56.7% vs. 50.0%). When using only vocal segments, classification accuracy increases to 64.6%. When using only music segments, accuracy reduces to 36.6%.

The best improvement of song accuracy relative to the baseline is 14% (64.9% vs. 56.7% accuracy), acheived by using voice segments on the 21-artist set with the dynamic-threshold voice-segmentation system. Using music-only segments on the same data set with the fixed threshold policy, the system performs 35% worse relative to baseline (36.6% vs. 56.7%).

Data set	baseline	vox-fixed	vox-dynamic	mus-fixed	mus-dynamic
21 artists	25.7 / 56.7	29.3 / 55.2	31.8 / 64.9	20.7 / 36.6	19.0 / 31.3
16 artists	30.3 / 65.3	34.7 / 58.4	35.6 / 68.3	24.8 / 47.5	22.0 / 44.6
5-matched	54.6 / 85.4	60.4 / 90.5	62.6 / 93.7	44.2 / 71.4	40.7 / 69.8
5-matched (optimized)	54.6 / 92.1	60.4 / 93.7	62.6 / 93.7	44.2 / 76.2	40.7 / 69.8
5-mismatched	37.2 / 54.5	37.3 / 54.5	39.6 / 50.9	34.9 / 49.1	34.5 / 50.9
5-mismatched (optimized)	37.2 / 58.2	37.3 / 60.0	39.6 / 61.8	34.9 / 58.2	34.5 / 52.7

Table 2: Artist classification accuracy (frame accuracy / song accuracy). The 16 artist set excludes 5 artists who do not use conventional vocals in their music. Sets 5-matched and 5-mismatched contain the same 2 albums from each of the same 5 artists, but 5-matched is trained and tested on songs from both albums, whereas 5-mismatched is trained on one album and tested on the other. The optimized rows present the best results over all variations of the summarization scoring algorithm, as described in section 2.6.

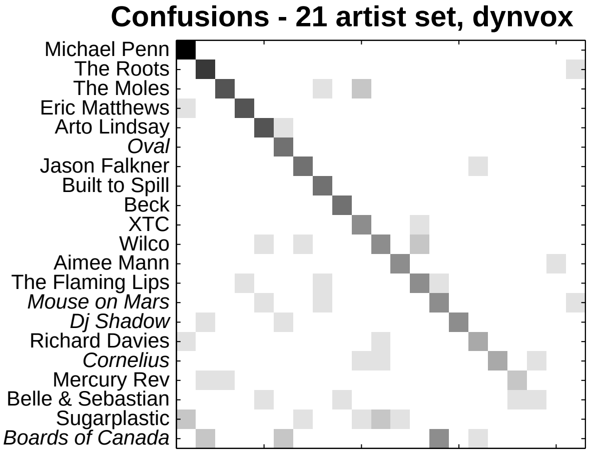

Figure 3: Confusion matrix for the 21 artist set, classified by the best-peforming vox-dynamic scheme. Artists are sorted in descending order of classification accuracy. Artists whose names appear in italics are the nonvocal musicians who were excluded from the 16 artist set.

The dynamic threshold does better than the fixed threshold; the average relative improvement over all test sets using the dynamic threshold is 6.7%. The improvement is not as significant on the smaller sets partly because the baseline already does well and there is not much room to grow. The system does 23% worse than the baseline when given music-only segments, averaged over all conditions and sets.

To summarize, using only voice helps somewhat, but using only non-voice (music) hurts much more. One interpretation of this result is that voice segments are generally more useful than music segments in terms of distinguishing artists, but when presented with the unsegmented data the classifier does fairly well at figuring this out on its own. However, when it is given only music it does not have the option to use voice, and performance suffers accordingly.

Figure 3 shows the confusion matrix for the vox-dynamic 21-artist case (our best classifier), with the artists sorted in descending order of classification accuracy; thus, Michael Penn was recognized with 100% accuracy, but none of the Sugarplastic or Boards of Canada tracks were recognized as such. Although this is a measure of the relative consistency between tracks from the classifier's point of view, there is no irresistable correlate in terms of subjective impression from listening to the music—certainly, the tracks on the Sugarplastic album cover a range of styles, but it would be very unfair to claim that all Michael Penn tracks sound exactly alike. Interestingly, some of the nonvocal artists (indicated by italicized names) were still classified quite accurately, such as Oval: this is consistent with the impression that any portion of their weirdly ambient music is easily recognized.

Note the performance difference between the fixed- and dynamic-threshold strategies: in all cases when judging by frame accuracy (and in all but one of the 5-artist cases judging by song-level accuracy), vox-dynamic outperforms vox-fixed, while in contrast music-fixed outperforms music-dynamic. This difference could be related to the amount of data resulting from the segmentation. For some songs with little or no vocals, the fixed threshold will select very few frames (less than $\frac{1}{2}$% of frames in the worst case, and below 5% of frames in 9 songs). This disadvantage does not hold in the music-segmentation case (only 1 song results in less than 5% of frames selected), and the fixed threshold on average segments more frames than the dynamic threshold (about 40% vs. about 25%).

The various summarization variants did not significantly alter the results for the 21- and 16-artist sets. For the 5-artist set, the best result over all variations of the summarization routine is also presented in Table 2. Note that the variant that produced the best result is not the same for all cases (but usually it was inverse-entropy with a minimum-confidence threshold), and moreover, the best result may not be chosen based on analysis of the training set. However, the optimized results provide an indication of the performance of each method under optimal selection of the summarization method (in practice, a separate cross-validation set could be used).

Due to the smaller number of artists and songs in the 5-artist cases, small differences in song-level accuracy may not being significant, however we can make the following observations. As expected, accuracy on the mismatched training and test sets is significantly lower than for the matched sets. Our results indicate that voice segments are better for classification than music segments, suggesting that the singer's voice is more consistent than the music across different songs from the same album. The mismatched dataset results also indicate that the singer's voice is more consistent across songs from different albums compared to the music, although the difference does not appear to be as great. An open question is whether the drastic reduction in performance seen in the mismatched set is primarily due to mismatched production quality, musical stylistic themes, instrumentation, or some other factor.

Figure 4 illustrates a typical case of the artist classifier in action. Each of the 21 posterior-probability outputs of the classifier neural net is show as a grayscale stripe against time, giving a spectrogram-like display which we term a 'posteriorgram'. In this case, vox-dynamic classification gives the correct result, as does classification based on the whole track; although there is some resemblence to Arto Lindsay in the first 30 seconds or so, the classification in this case is reasonably unambiguous.

Figure 5 provides a closer look at an example where using voice segments corrects an erroneous classification at the unsegmented level. Note that the automatically-identified voice segments (shaded, outlined boxes) are not particularly accurate compared to the hand-marked vocal segments (shown immediately above the main plot) partly because the dynamic threshold forces 25% of the frames to be labeled voice, far more than are actually voice in this example, and partly because the vocals are spoken softly in this track, rather than sung prominently as in the bulk of the training set. Interestingly, although the segmentation is not accurate, it manages to choose frames which are correctly recognized, and discards frames that are incorrectly identified as "Oval". The track consists mostly of a slow, bassy acoustic guitar line that does not sound unlike Oval's throbbing ambient electronica, whereas most of the other Arto Lindsay tracks in the training set contain drums.

On the other hand, Figure 6 illustrates a case where the voice segments are deceptive. In fact there are not really any vocals in this track by Cornelius, it is a cut-up collage of samples taken from old television and movie themes, set to a William Tell-speed drum loop. There are some vocal samples mixed in, but they consist of different voices, none of them the artist's. In this case, it is not surprising that the music segments are more correct. This example was correctly identified by the unsegmented system, but mistakenly identified as XTC when only the "vocal" segments are used. Once again, the dynamic threshold policy forces the system to select many poor choices as vocal segments.

4. FUTURE WORK

Many questions remain. A notable few are:

- How does the performance of the vocals segment detector affect the results? One way to shed light on this question would be to perform hand-labeling of voice segments for our entire 269 song set, to see how differently the true voice segments would perform.

- What features are better than MFCCs for singer discrimination? An initial experiment with using time-varying statistics of MFCCs showed promise. Would other features be more appropriate for capturing vocals than music? Although MFCCs have been developed primarily for speech recognition, they have on many occasions proved useful in music-related tasks, for instance [6, 5].

- Is the "album effect" primarily due to audio production techniques, musical stylistic elements, instrumentation, or some other factor? An experiment can be devised to control for some of these possible causes by, for example, choosing a set of multiple distinct recordings of the same songs by the same artists. If production effects turn out to be dominant, can they be reduced by simple feature normalization such as cepstral mean subtraction? Can the album effect be reduced by simply training on a more diverse data set?

- What other information can be extracted from the posteriograms? Perhaps they can tell us something about the song structure, or perhaps they can be used to cluster artists with similar voices for use in a recommendation engine.

- We are working on a method to roughly separate voice from music based on the stereo center-channel subtraction method well-known in karaoke applications. One variant can be inverted to obtain the voice isolated from the music. Even simple band-pass filtering could exclude much of the nonvocal energy due to bass and drums. How does the artist classifier perform when trained on a voice signal isolated in this way?

5. CONCLUSION

We presented a system for classifying musical tracks according to artist that first identified the most vocals-like

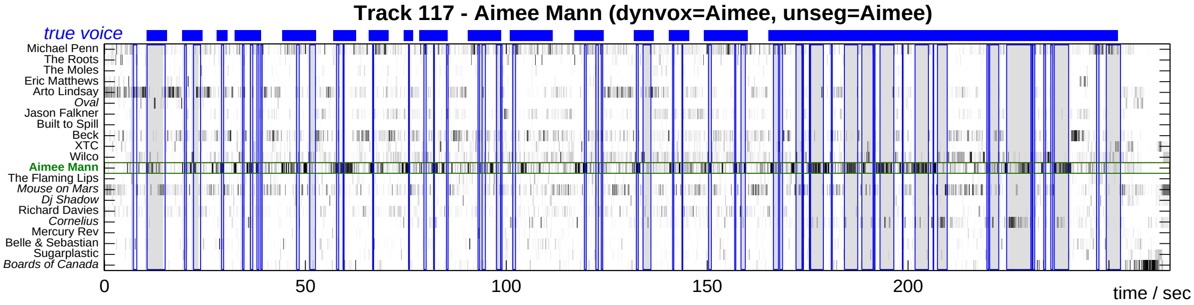

Figure 4: Illustration of the artist classifier output. Time goes from left to right; this track is 265 seconds long. Rows display the instantaneous posterior probability estimates for each of the 21 artists. Outlined, shaded regions in the main panel show the frames selected as vocal-like by the automatic voice detector. These can be compared to the hand-marked vocal segments shown by the thick lines above the main panel. This song contains a large proportion of singing, and almost all of the automatically-detected segments occur within true vocal segments. The dynamic threshold policy has restricted the segmentation to include only the 25% of frames that are most 'voice-like'. The correct vox-dynamic classification of this track as Aimee Mann is indicated by the horizontal green box and the highlighted name on the left. This classification was based only on the shaded vocals-tagged segments, although in this case using the entire, unsegmented track would give the same classification result.

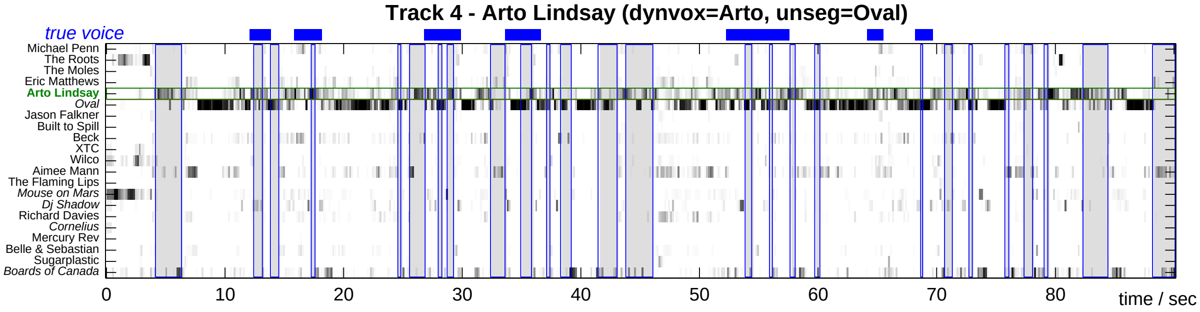

Figure 5: Posteriogram for an example that benefited by using only vocal segments, correctly classifying the track as Arto Lindsay, when classifying on the entire, unsegmented track returned the incorrect assignment of Oval. It can be seen that Oval is given high posterior probability over much of the track, but mainly in the regions not labeled as voice, which are ignored in the vox-dynamic classification scheme.

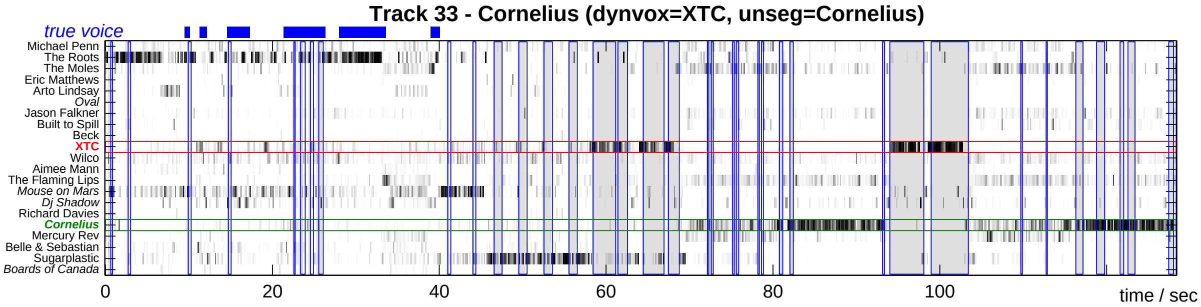

Figure 6: Posteriogram for an example where voice segments are deceptive. This Cornelius track is correctly classified as such when the unsegmented data is used, but within the most vocal-like segments (most of which in fact contain no voice), the dominant classification is as XTC, shown by the red outline.

segments within the track, then performed classification based only on those segments. For a medium-sized database with good consistency within artists, constructed from a single album for each of 21 artists, the best classification accuracy achieved was 64.9%, which is to our knowledge the best result reported for such a task, and is significantly better than equivalent classification based only on segments labeled as nonvocal, or on the full, unsegmented audio. However, in a small test of cross-album generalization (two albums for each of five artists), performance was much less impressive, and the preferential use of vocal segments offered little or no advantage. Further work is needed to devise features less sensitive to the "album effects" that differentiate individual albums, in addition to several other open questions raised by this study.

ACKNOWLEDGMENTS

Our thanks go to Brian Whitman for helpful discussions and for sharing his data. This work was funded by NEC Research Institute whose support is very gratefully acknowledged.

REFERENCES

[1] E.D. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1331–1334, 1997.

[2] Adam Berenzweig and Daniel P.W. Ellis. Locating singing voice segments within music signals. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001.

[3] Dominic Genoud, Dan Ellis, and Nelson Morgan. Combined speech and speaker recognition with speaker-adapted connectionist models. In *Automatic Speech Recognition and Understanding Workshop*, Keystone, CO, December 12–15 1999.

[4] D. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Trans. Speech and Audio Processing*, 2(4):639–643, 1994.

[5] Beth Logan. Mel frequency cepstral coefficients for music modelling. In *International Symposium on Music Information Retrieval*, Plymouth, MA, October 23–25 2000.

[6] Brian Whitman, Gary Flake, and Steve Lawrence. Artist detection in music with minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568. Falmouth, Massachusetts, September 10–12 2001.