# LOCATING SINGING VOICE SEGMENTS WITHIN MUSIC SIGNALS

*Adam L. Berenzweig and Daniel P.W. Ellis*

Dept. of Electrical Engineering, Columbia University, New York 10027
alb63@columbia.edu, dpwe@ee.columbia.edu

## ABSTRACT

A sung vocal line is the prominent feature of much popular music. It would be useful to reliably locate the portions of a musical track during which the vocals are present, both as a 'signature' of the piece and as a precursor to automatic recognition of lyrics. Here, we approach this problem by using the acoustic classifier of a speech recognizer as a detector for speech-like sounds. Although singing (including a musical background) is a relatively poor match to an acoustic model trained on normal speech, we propose various statistics of the classifier's output in order to discriminate singing from instrumental accompaniment. A simple HMM allows us to find a best labeling sequence for this uncertain data. On a test set of forty 15 second excerpts of randomly-selected music, our classifier achieved around 80% classification accuracy at the frame level. The utility of different features, and our plans for eventual lyrics recognition, are discussed.

## 1. INTRODUCTION

Popular music is fast becoming one of the most important data types carried by the Internet, yet our ability to make automatic analyses of its content is rudimentary. Of the many kinds of information that could be extracted from music signals, we are particularly interested in the vocal line i.e. the singing: this is often the most important 'instrument' in the piece, carrying both melodic 'hooks' and of course the lyrics (word transcript) of the piece. It would be very useful to be able to transcribe song lyrics with an automatic speech recognizer, but this is currently impractical: singing differs from speech in many ways, including the phonetic and timing modifications employed by singers, the interference caused by the instrumental background, and perhaps even the peculiar word sequences used in lyrics. However, as a first step in the direction of lyrics recognition, we are studying the problem of locating the segments containing voice from within the entire recording, i.e. building a 'singing detector' that can locate the stretches of voice against the instrumental background.

Such a segmentation has a variety of uses. In general, any kind of higher-level information can support more intelligent handling of the media content, for instance by automatically selecting or jumping between segments in a sound editor application. Vocals are often very prominent in a piece of music, and we may be able to detect them quite robustly by leveraging knowledge from speech recognition. In this case, the pattern of singing within a piece could form a useful 'signature' of the piece as a whole, and one that might robustly survive filtering, equalization, and digital-analog-digital transformations.

Transcription of lyrics would of course provide very useful information for music retrieval (i.e. query-by-lyric) and for grouping different versions of the same song. Locating the vocal segments within music supports this goal at recognition-time, by indicating which parts of the signal deserve to have recognition applied. More significantly, however, robust singing detection would support the development of a phonetically-labeled database of singing examples, by constraining a forced-alignment between known lyrics and the music signal to search only within each phrase or line of the vocals, greatly improving the likely accuracy of such an alignment.

Note that we are assuming that the signal is known to consist only of music, and that the problem is locating the singing within it. We are not directly concerned with the problem of distinguishing between music and regular speech (although our work is based upon these ideas), nor the interesting problems of distinguishing vocal music from speech [1] or voice-over-music from singing—although we note in passing that the approach to be described in section 2 could probably be applied to those tasks as well.

The related task of speech-music discrimination has been pursued using a variety of techniques and features. In [2], Scheirer and Slaney defined a large selection of signal-level features that might discriminate between regular speech and music (with or without vocals), and reported an error rate of 1.4% in classifying short segments from a database of randomly-recorded radio broadcasts as speech or music. In [3], Williams and Ellis attempted the same task on the same data, achieving essentially the same accuracy. However, rather than using purpose-defined features, they calculated some simple statistics on the output of the acoustic model of a speech recognizer (a neural net estimating the posterior probability of 50 or so linguistic categories) applied to the segment to be classified; since the model is trained to make fine distinctions among speech sounds, it responds very differently to speech, which exhibits those distinctions, as compared to music and other nonspeech signals that rarely contain 'good' examples of the phonetic classes.

Note that in [2] and [3], the data was assumed to be presegmented so that the task was simply to classify predefined segments. More commonly, sound is encountered as a continuous stream that must be segmented as well as classified. When dealing with pre-defined classes (for instance, music, speech and silence), a hidden Markov model (HMM) is often employed (as in [4]) to make simultaneous segmentation and classification.

The next section presents our approach to detecting segments of singing. Section 3 describes some of the specific statistics we tried as a basis for this segmentation, along with the results. These results are discussed in section 4, then section 5 mentions some ideas for future work toward lyric recognition. We state our conclusions in section 6.

## 2. APPROACH

In this work, we apply the approach of [3] of using a speech recognizer's classifier to distinguishing vocal segments from accompaniment: Although, as discussed above, singing is quite different from normal speech, we investigated the idea that a speech-trained acoustic model would respond in a detectably different manner to singing (which shares some attributes of regular speech, such as formant structure and phone transitions) than to other instruments.

We use a neural network acoustic model, trained to discriminate between context-independent phone classes of natural English speech, to generate a vector of posterior probability features (PPFs) which we use as the basis for our further calculations. Some examples appear in figure 1, which shows the PPFs as a 'posteriogram', a spectrogram-like plot of the posterior probability of each possible phone-class as a function of time. For well-matching natural speech, the posteriogram is characterized by a strong reaction to a single phone per frame, a brief stay in each phone, and abrupt transitions from phone to phone. Regions of non-speech usually show a less emphatic reaction to several phones at once, since the correct classification is uncertain. In other cases, regions of non-speech may evoke a strong probability of the 'background' class, which has typically been trained to respond to silence, noise and even background music. Alternatively, music may resemble certain phones, causing either weak, relatively static bands or rhythmic repetition of these "false" phones in the posteriogram.

Within music, the resemblance between the singing voice and natural speech will tend to shift the behavior of the PPFs closer toward the characteristics of natural speech when compared to non-vocal instrumentation, as seen in figure 1. The basis of the segmentation scheme presented here is to detect this characteristic shift. We explore three broad feature sets for this detection: (1) direct modeling of the basic PPF features, or selected class posteriors; (2) modeling of derived statistics, such as classifier entropy, that should emphasize the differences in behavior of vocal and instrumental sound; and (3) averages of these values, exploiting the fact that the timescale of change in singing activity is rather longer than the phonetic changes that the PPFs were originally intended to reveal, and thus the noise robustness afforded by some smoothing along the time axis can be usefully applied.

The specific features investigated are as follows:

- 12th order PLP cepstral coefficients plus deltas and double-deltas. As a baseline, we tried the same features used by the neural net as direct indicators of voice vs. instruments.
- Full log-PPF vector i.e. a 54 dimensional vector for each time frame containing the pre-nonlinearity activations of the output layer of the neural network, approximately the logs of the posterior probabilities of each phone class.
- Likelihoods of the log-PPFs under 'singing' and 'instrument' classes. For simplicity of combination with other unidimensional statistics, we calculated the likelihoods of the 54-dimensional vectors under the multidimensional full-covariance Gaussians derived from the singing and instrumental training examples, and used the logs of these two likelihoods, PPF $L_{voc}$ and $L_{mus}$, for subsequent modeling.
- Likelihoods of the cepstral coefficients under the two classes. As above, the 39-dimensional cepstral coefficients are evaluated under single Gaussian models of the two classes to produce Cep $L_{voc}$ and $L_{mus}$.
- Background log-probability $log(P_{bg})$. Since the background class has been trained to respond to nonspeech, and since its value is one minus the sum of the probability of all the actual speech classes, this single output of the classifier is a useful indicator of voice presence or absence.
- Classifier entropy. Following [3], we calculate the per-frame entropy of the posterior probabilities, defined as:

$$H(n) = \sum_k -p(q_k^n) log(p(q_k^n)) \quad (1)$$

  where $q_k^n$ is the posterior probability of phone class $k$ at time $n$. This value should be low when the classifier is confident that the sound belongs to a particular phone class (suggesting that the signal is very speech-like), or larger when the classification is ambiguous (e.g. for music).

  To separate the effect of a low entropy due to a confident classification as background, we also calculated the entropy-excluding-background $H_{\overline{bg}}$ as the entropy over the 53 true phonetic classes, renormalized to sum to 1.
- Dynamism. Another feature defined in [3] is the average sum-squared difference between temporally adjacent PPFs i.e.

$$D(n) = \sum_k (p(q_k^n) - p(q_k^{n-1}))^2 \quad (2)$$

  Since well-matching speech causes rapid transitions in phone posteriors, this is larger for speech than for other sounds.

Because our task was not simply classification of segments as singing or instrumental, but also to make the segmentation of a continuous music stream, we used an HMM framework with two states, "singing" and "not singing", to recover a labeling for the stream. In each case, distributions for the particular features being used were derived from hand-labeled training examples of singing and instrumental music, by fitting a single multi-dimensional Gaussian for each class to the relevant training examples. Transition probabilities for the HMM were set to match the label behavior in the training examples (i.e. the exit probability of each state is the inverse of the average duration of segments labeled with that state).

## 3. RESULTS

### 3.1. Speech model

To generate the PPFs at the basis of our segmentation, we used a multi-layer perceptron neural network with 2000 hidden units, trained on the NIST Broadcast News data set to discriminate between 54 context-independent phone classes (a subset of the TIMIT phones) [5]. This net is the same as used in [3], and is publicly available. The net operates on 16 ms frames i.e. one PPF frame is generated for each 16 ms segment of the data.

### 3.2. Audio data

Our results are based on the same database used in [2, 3] of 246 15-second fragments recorded at random from FM radio in 1996. Discarding any examples that do not consist entirely of (vocal or

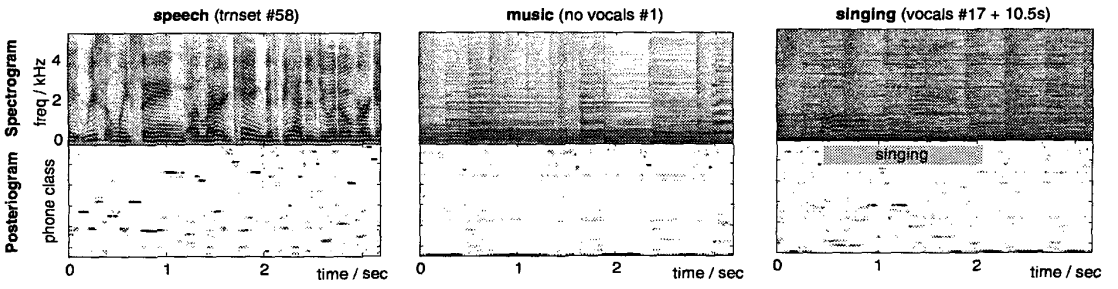


Figure 1: Spectrograms of a speech example and two musical fragments, with and without singing, along with the 'posteriograms' showing the output of the speech classifier. The singing in the rightmost example (marked by the gray bar) evokes a distinctive pattern of response in the posteriogram.

instrumental) music leaves 101 fragments, 66 of which contain vocals. We hand-labeled the vocals examples to mark the precise segments containing singing; typically one sung phrase would result in a single segment. The average duration of singing segments was 5.5 seconds. 40 fragments were randomly selected as a test set. The remaining 61 fragments were used as labeled training data.

### 3.3. Scoring

Table 1 shows the performance of segmentation based on various statistics and combinations. The results are given as frame error rate, i.e. for every 16 ms frame in the test data, the label assigned by the best path through the HMM (based on the statistics or features shown) is compared to the 'ground truth' label from the hand-marking. This measure does not differentiate between errors due to boundaries that are shifted in time and errors due to inserted or deleted segments of singing (both kinds of errors occurred). However, the frame error rate provides a reasonable relative performance measure.

For each feature basis, the results of averaging the features over different numbers of frames are shown (where averaging over one frame is just using the features directly). The 16 ms frame resolution of the speech classifier was much finer than needed for the segmentation task, and averaging over a longer time window helped smooth out frame-to-frame variations to reveal the underlying trends.

These results are also plotted in figure 2, which shows the variation of frame error rate for several different feature bases as a function of averaging window length (for a wider range of windows than reported in table 1). We see that averaging improves performance fairly uniformly out to 81 frames (1.3 seconds), but beyond that, the averaging window is longer than many of the segments to be detected, and performance begins to decline. In each case, the HMM is finding labels for each 16 ms frame, although a practical system would use a coarser resolution.

| Features/stats | Classification Frame Error Rate 1 frame | 9 frame | 81 frame |
|---|---|---|---|
| 39 Cepstra | 31.4% | 26.3% | 29.4% |
| 54 log-PPFs | 35.2% | 31.0% | 31.2% |
| Cep $log(L_{mus})$ & $log(L_{voc})$ | 35.2% | 31.0% | 31.2% |
| PPF $log(L_{mus})$ & $log(L_{voc})$ | 25.1% | 23.5% | 20.4% |
| $log(P_{bg})$ | 41.3% | 40.6% | 40.3% |
| Entropy $H$ | 38.6% | 36.2% | 36.1% |
| $H_{\overline{bg}}$ | 35.5% | 36.4% | 35.8% |
| Dynamism $D$ | 44.8% | 44.8% | 44.6% |
| All 6 stats | 28.8% | 29.0% | 21.3% |
| Best 3 | 26.1% | 26.6% | 18.8% |

Table 1: Frame error rate for vocals/instrumental segmentation based on different features or statistics, either using the values at each frame ("1 frame"), or averaging the features within overlapping windows of 9 or 81 frames. "All 6 stats" refers to the combination of the four individual statistics shown in the third panel plus the PPF-based $log(L_{mus})$ and $log(L_{voc})$. "Best 3" refers to the best-performing combination of PPF $log(L_{mus})$ and $log(L_{voc})$ combined with $H$.

## 4. DISCUSSION

It is disappointing that our carefully-designed custom statistics performed no better than direct modeling of the raw high dimensional feature space, and indeed that the raw PPFs produced by the neural network classifier gave more errors than the raw cepstral coefficients. However, the PPF-based likelihoods $L_{voc}$ and $L_{mus}$ do outperform the cepstral baseline, especially in combination with one of the hand-designed features such as entropy $H$.

We note the significant improvement achieved by adding a further stage of simple Gaussian modeling on the 2-D feature space formed by the log-likelihoods PPF $log(L_{mus})$ and $log(L_{voc})$ (obtained from the 54 dimensional baseline Gaussian models). Since there is basically no additional information available at this second stage of calculation, this indicates a modeling weakness: we could presumably match or better this result e.g. by using Gaussian mixture models (GMMs) in the original high-dimensional space.

The cepstral-based features did not improve with time-averaging over a window longer than 9 frames. Presumably, the rapid rate of change of the cepstrum leads to within-class variation that is too great to be amenable to a longer smoothing window. The fact that the PPF-based features do improve with longer time averaging confirms that they don't use fine temporal structure of phone transitions such as our hand-designed features were designed to detect, but rather characterize the overall distribution of phones.

In browsing the labeling errors, we saw many instances of short excursions into the incorrect class, particularly when the av-

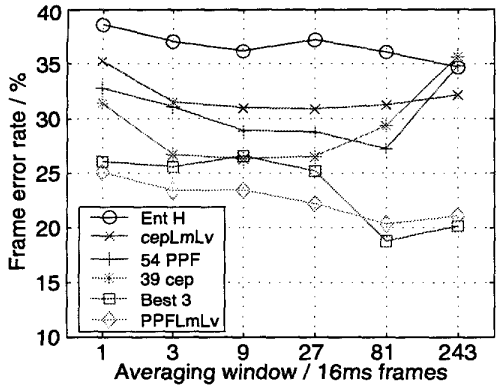


Figure 2: Variation of vocals/accompaniment labeling frame error rate as a function of averaging window length in frames (each frame is 16 ms, so a 243 frame window spans 3.9 sec).

eraging window was short. Imposing a minimum label duration of several hundred milliseconds would not exclude any of the ground-truth segments, so these errors could be eliminated with slightly more complicated HMM structure that enforces such a minimum duration through repeated states.

What began as a search for a few key features has led to a high-order, but more task-independent, modeling solution: In [2], a number of unidimensional functions of an audio signal were defined that should help to distinguish speech from music, and good discrimination was achieved by using just a few of them. In [3], consideration of the behavior of a speech recognizer's acoustic model similarly led to a small number of statistics which were also sufficient for good discrimination. In the current work, we attempted a related task—distinguishing singing from accompaniment—using similar techniques. However, we discovered that training a simple high-dimensional Gaussian classifier directly on speech model outputs—or even on the raw cepstra—performed as well or better.

At this point, the system resembles the 'tandem acoustic models' (PPFs used as inputs to a Gaussian-mixture-model recognizer) that we have recently been using for speech recognition [6]. Our best performing singing segmenter is a tandem connection of a neural-net discriminatory speech model, followed by a high-dimensional Gaussian distribution model for each of the two classes, followed by another pair of Gaussian models in the resulting low-dimensional log-likelihood space. One interpretation of this work is that it is more successful, when dealing with a reasonable quantity of training data, to train large models with lots of parameters and few preconceptions, than to try to 'shortcut' the process by defining low-dimensional statistics. This lesson has been repeated many times in pattern recognition, but we still try to better it by clever feature definitions.

## 5. FUTURE WORK

As discussed in the introduction, this work is oriented toward the transcription of lyrics as a basis for music indexing and retrieval. It is clear (e.g. from figure 1) that using a classifier trained on normal speech is too poorly matched to the acoustics of singing in popular music to be able to support accurate word transcription. More promising would be a classifier trained on examples of singing. To obtain this, we need a training set of singing examples aligned to their lexical (and ultimately phonetic) transcriptions. The basic word transcripts of many songs— i.e. the lyrics—are already available, and the good segmentation results reported here provide the basis for a high-quality forced alignment between the music and the lyrics, at least for some examples, even with the poorly-matched classifier.

Ultimately, however, we expect that in order to avoid the negative effect of the accompanying instruments on recognition, we need to use features that can go some way toward separating the singing signal from other sounds. We see Computational Auditory Scene Analysis, coupled with Missing-Data speech recognition and Multi-Source decoding, as a very promising approach to this problem [7].

## 6. CONCLUSIONS

We have focused on the problem of identifying segments of singing within popular music as a useful and tractable form of content analysis for music, particularly as a precursor to automatic transcription of lyrics. Using Posterior Probability Features obtained from the acoustic classifier of a general-purpose speech recognizer, we were able to derive a variety of statistics and models which allowed us to train a successful vocals detection system that was around 80% accurate at the frame level. This segmentation is useful in its own right, but also provides us with a good foundation upon which to build a training set of transcribed sung material, to be used in more detailed analysis and transription of singing.

## 7. ACKNOWLEDGMENTS

We are grateful to Eric Scheirer, Malcolm Slaney and Interval Research Corporation for making available to us their database of speech/music examples.

## 8. REFERENCES

[1] W. Chou and L. Gi, Robust singing detection in speech/music discriminator design," *Proc. ICASSP*, Salt Lake, May 2001

[2] E. Scheirer and M. Slaney "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP*, Munich, April 1997.

[3] G. Williams and D. Ellis "Speech/music discrimination based on posterior probability features," *Proc. Eurospeech*, Budapest, September 1999.

[4] T. Hain, S. Johnson, A. Tuerk, P. Woodland and S. Young, "Segment Generation and Clustering in the HTK Broadcast News Transcription System," *Proc. DARPA Broadcast News Workshop*, Lansdown VA, February 1998.

[5] G. Cook et al., "The SPRACH System for the Transcription of Broadcast News," *Proc. DARPA Broadcast News Workshop*, February 1999.

[6] H, Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proc. ICASSP*, Istanbul, June 2000.

[7] J. Barker, M. Cooke and D. Ellis, "Decoding speech in the presence of other sound sources," *Proc. ICSLP*, Beijing, October 2000.