SOLO VOICE DETECTION VIA OPTIMAL CANCELLATION

Christine Smit and Daniel P.W. Ellis[*]

LabROSA, Electrical Engineering
Columbia University
New York NY 10025 USA
`{csmit,dpwe}@ee.columbia.edu`

ABSTRACT

Automatically identifying sections of solo voices or instruments within a large corpus of music recordings would be useful, for example, to construct a library of isolated instruments to train signal models. We consider several ways to identify these sections, including a baseline classifier trained on conventional speech features. Our best results, achieving frame level precision and recall of around 70%, come from an approach that attempts to track the local periodicity of an assumed solo musical voice, then classifies the segment as a genuine solo or not on the basis of what proportion of the energy can be canceled by a comb filter constructed to remove just that periodicity.

1. INTRODUCTION

This work is motivated by a project to model the statistics of professional singers' voices, for which we would like to assemble a large collection of solo voice recordings. Many existing music recordings contain some solo voice, but manually marking the solo passages would severely limit the amount of data we can obtain. An automatic system for identifying stretches of solo voice would allow us to mine large online music audio archives to obtain essentially unlimited quantities of solo voice or other solo instruments. Finding these "unobstructed" views of musical instruments is valuable for many applications of modeling single voices, such as to be able to recognize them better in the context of other instruments (e.g. [1]).

Our approach is based on the idea that a solo musical passage will for the most part consist of a single note (pitch) sounding at any time. The spectral structure of an isolated pitch is characteristically simple, consisting of well-defined, regularly spaced harmonic spectral peaks(as illustrated in the top pane of figure 1) and this should allow us to distinguish these frames from either multiple simultaneous voices (middle pane) which exhibit a much more complex pattern of superimposed and interacting harmonic series, or silent gaps (bottom pane) which reveal a frequency-dependent noise floor.

We considered several different approaches. Our baseline system adopts the same approach used for detecting when a singer is active during accompanied music [2] by training a classifier on the ubiquitous Mel-frequency cepstral coefficients (MFCCs) borrowed from speech recognition.

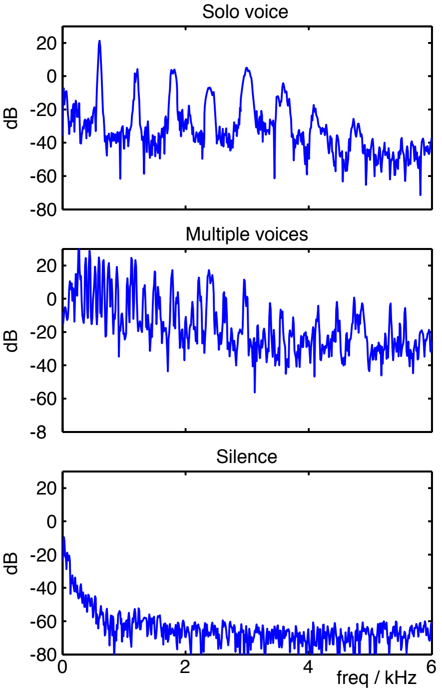


Figure 1: Example spectra taken from solo voice (top pane), ensemble accompaniment (middle pane), and background silence (bottom pane).

We were also interested in seeing if the specific structural details visible in figure 1 could be employed directly. Our first idea was to attempt to spot the 'gaps' between the harmonics of the solo voice which might be expected to revert to the noise floor. However, we found it difficult to make this approach work, particularly as the voice pitch became lower and the 'gaps' became smaller.

Our most successful approach is based on the idea of attempting to model a short-time frame of the signal as consisting of a single periodicity, canceling energy at that period with an appropriate comb filter (i.e. subtracting the signal delayed by one period from itself), then seeing what proportion of the total signal energy is removed. When the signal consists largely or wholly of a single periodicity, it should be possible to cancel virtually all of the periodic (tonal) energy, leading to a very large drop in energy after the filter.

---

[*] This work was supported by the Columbia Academic Quality Fund, and by the National Science Foundation (NSF) via a fellowship and Grants No. IIS-0238301 and IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

In general, however, the optimal period will not be an integer number of samples, so a fractional-delay filter is required. The next section describes our approach to finding this filter, then section 3 describes our experiments with this detector, comparing it to our MFCC-based baseline. Section 4 concludes with a discussion of the filter and how it works in practice, and considers some other applications.

2. OPTIMAL PERIODICITY CANCELLATION

By definition, a single voice has a single pitch (in the sense of a fundamental frequency), which, for musical voices, will often be relatively stationary. To detect if only a single voice is present, our approach is to find the best-fitting single period, cancel its energy, and see how completely this has removed the energy of the frame. Solo voices will have only their aperiodic energy left, resulting in a large drop in energy. Polyphonies consisting of several instruments playing different pitches will have canceled only one of the periodicities, leading to a much smaller drop in energy.

After breaking up our soundfiles into 93 ms frames (i.e. 4096 samples at 44.1 kHz sampling rate), we use autocorrelation to obtain an initial estimate, τ, of the dominant fundamental period for each frame by finding the largest peak in the autocorrelation over all lags. A simple filter (figure 2, top) might then be able to remove that frequency and all its harmonics:

$$\epsilon[n] = x[n] - x[n-\tau] \tag{1}$$

If τ exactly matches the period of a purely periodic waveform within the frame, $\epsilon[n]$ should be identically zero.

The problem with this scheme is that, in general, the period of an acoustic source will not correspond to an integer number of samples. This problem has been encountered in many previous circumstances including the "long-term predictor" of traditional vocoders [3] and the delay lines at the heart of physical modeling music synthesis [4]. To get around this limitation, we employ a slightly more complicated filter to optimally remove the voice (figure 2, bottom),

$$\epsilon[n] = x[n] - \sum_{i=-k}^{k} a[i] \cdot x[n - (\tau + i)] \tag{2}$$

or

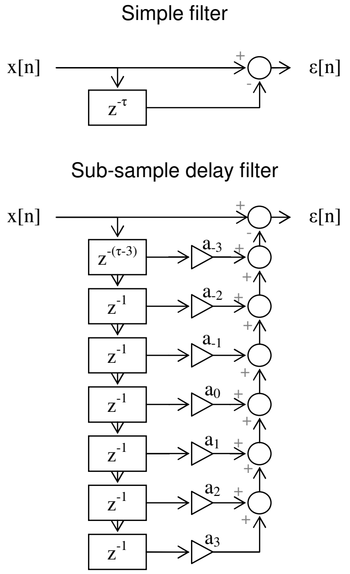
$$\mathbf{e} = \mathbf{x} - \mathbf{Z}\mathbf{a} \tag{3}$$

where $\mathbf{e}_i = \epsilon[i]$, $\mathbf{x}_i = x[i]$, $\mathbf{Z}_{i,j} = x[i - (\tau + j)]$, $\mathbf{a}_j = a[j]$; $i \in [0, N-1]$ and $j \in [-k, k]$. We used $k = 3$ for a seven-coefficient filter as a more or less arbitrary compromise between computational complexity and flexibility of the cancellation filter.

The $a[i]$ coefficients that optimally reduce the energy of $\epsilon[n]$ are found by the least squares solution,

$$\hat{\mathbf{a}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{x} \tag{4}$$

Having solved for these coefficients within each frame, we apply the filter to find the energy of the residual $\epsilon[n]$ within the frame, then calculate the ratio of the residual energy to the energy of the original signal $x[n]$. In the case of a purely periodic signal whose period is a non-integer number of samples, we expect $\hat{\mathbf{a}}$ to approximate an ideal fractional delay filter (sinc interpolator) which can exactly cancel the periodic signal, leading to a residual-to-original ratio close to zero. When the signal consists of many periodicities, only a small proportion of the energy will be canceled by eliminating just one dominant periodicity.


Simple filter

Sub-sample delay filter

Figure 2: Optimal cancellation filters. Top: for signals with integer periodicities; bottom: a filter able to cancel non-integer periodicities.

In frames consisting of "silence" (noise floor), however, a single spectral peak may account for a large proportion of the very small amount of energy. In this case, the optimal cancellation filter may also be able to remove a large proportion of the energy. To differentiate between silent frames and single voice frames, we added a value related to each frame's original energy as a second feature. To avoid any issues arising from global scaling of the original sound files, we normalized the entire waveform to make the 98th percentile of the short-time Fourier transform magnitude equal to 1.

2.1. Classifier

We use the residual-to-original energy ratio and the normalized absolute energy as a two-dimensional feature and feed them to a simple Bayesian classifier to estimate the probability that each frame belongs to each of three classes – solo voice, multiple voices, and silence. We model the distribution of the features for each of these classes separately using a small amount of hand-labeled training data (see section 3). The normalized absolute energy is fit with a Gaussian in the log (dB) domain. The residual-to-original energy ratio, however, always lies between 0 and 1, and is heavily skewed toward 0 in the solo class. A Gaussian is thus a poor fit, and no simple transformation will make all the classes appear Gaussian. Instead, we model it with a Beta distribution for each category. The Beta distribution is defined over $[0, 1]$ and has two parameters to fit both the mode and spread of the observed class-conditional values. We treat the two features as independent, so we obtain the overall likelihood of a particular observation frame under each class by multiplying the Gaussian and Beta likelihoods together for

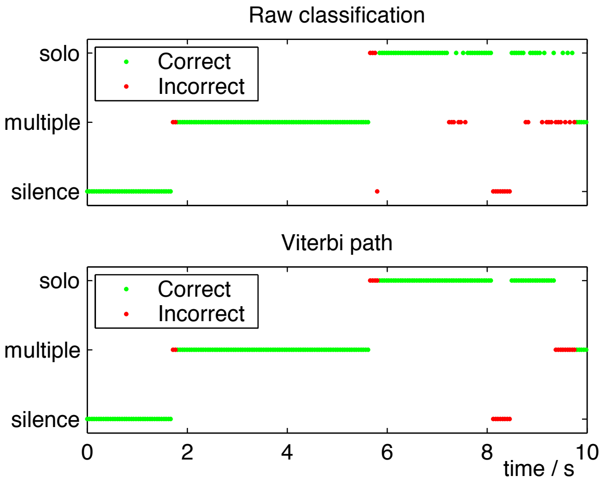


Figure 3: Example output from the cancellation-based classifier. Top pane shows raw frame-level results, bottom pane shows the result of HMM smoothing to remove rapid switching between states.

that class. Simple ML classification then is used to label according to the largest posterior.

Independent classification of each time frame can result in rapid alternation between class labels, whereas real data changes state relatively infrequently. We build a simple three-state hidden Markov model (HMM) with transition probabilities set to match the empirical frame transition counts in the training data. We can then find the single most likely label sequence given this transition model and the class-dependent likelihoods with the Viterbi algorithm. (We used Kevin Murphy's Matlab implementation [5].) Figure 3 shows an example of label sequences before and after HMM smoothing, compared to the ground-truth labels.

To trade precision for recall, we can bias the model to generate more or fewer "solo" labels simply by scaling the solo model likelihood by a constant value. Smaller likelihoods for the "solo" class result in fewer, more confidently "solo" labels. In our application, assuming a very large underlying archive to search, we might be happy to accept a low recall (only a small portion of all possible solo regions are identified) in order to achieve a higher precision (nearly all of the identified regions are, in fact, solo regions).

3. EXPERIMENTS

3.1. Data

Our data set consisted of twenty 1 minutes samples that were hand-labeled as silence, solo, or multiple voices. The samples were taken from a variety of professional folk and classical recordings. About 28% of the frames in the data set contained a solo voice. The other 72% of data frames contained multiple voices or silence. Ten samples were used calculating the distribution and Viterbi path parameters. The remaining ten samples were used for testing.

3.2. Baseline Classifier

As mentioned in the introduction, we also implemented a 'generic' audio classifier based on the Mel-frequency cepstral coefficient feature vectors commonly used in speech recognition and that have also shown themselves very successful in many music classification tasks [6, 2]. We used the first 13 cepstral coefficients and normalized their means to be zero within each track to eliminate any fixed filtering effects. The same Bayesian classifier structure was used, but in this case each of the three classes was fit by a full-covariance multidimensional Gaussian. We used Netlab for this modeling [7].

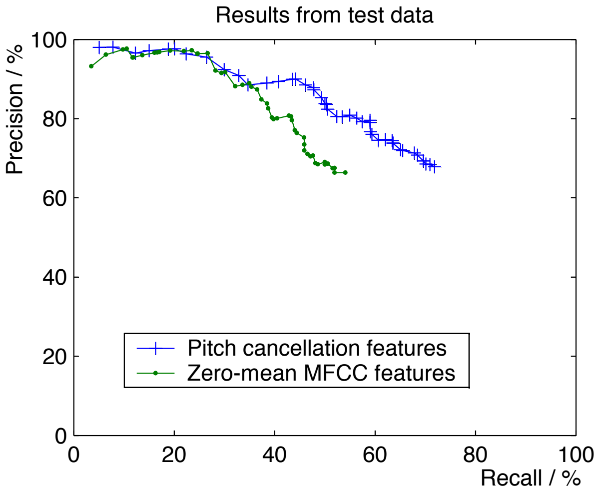


Figure 4: Solo voice detection precision/recall trade-off for the two approaches.

3.3. Results

Figure 4 shows the results of the cancellation- and MFCC-based classifiers on all the test data combined. We obtained different precision/recall points by manipulating the solo likelihood. Above about 90% precision, the MFCC and cancellation systems perform approximately equally. At lower precision levels, however, the cancellation algorithm has a much better recall. At 80% precision and below, the cancellation algorithm has at least 10% higher recall than the MFCC system.

The cancellation system also exhibits more consistent performance. When comparing the frame labeling accuracy on individual tracks in the test set, the variance of the cancellation system performance was half that of the MFCC system. We suspect this is because the pitch cancellation algorithm has many fewer learned parameters (4 per class, compared to 104 per class for the MFCC) and thus is less susceptible to overfitting.

4. DISCUSSION AND CONCLUSIONS

Although we resorted to the least-squares optimal FIR filter simply as a way to achieve precise fitting of non-integer-period signals, it is interesting to consider what we get as a result. The filter is optimized to minimize output energy, subject to the constraints that (a) the first value of the impulse response is 1; (b) the next $\tau - 4$ are zero; and (c) the total length is $\tau + 3$, where τ is our initial, coarse estimate of the dominant period. This filter is not constrained to include an exact unit-gain, linear-phase fractional delay, and in general energy will be minimized by a more complex response subject to the constraints. The seven free parameters of our system allow a certain error tolerance in the coarse period estimation as well as making it possible to match an ideal sync fractional delay more accurately, but they also permit a more complex range of solutions; solving for longer filter blocks would result in filters that deviate increasingly far from our intention of a simple comb canceling a single period.

Results from running the optimal cancellation algorithm on a partial music track can be seen in figure 5. Silence starts the track and then the full orchestra enters(at around 2 seconds). The orchestra drops out (around 6 seconds) for the soloist's entrance and then joins back in the mix (around 10 seconds). While the orchestra is playing, the spectrum is quite dense and the comb filter cannot adapt to remove energy effectively. As soon as the soloist enters, however, the filter takes advantage of the harmonic structure to remove a significant portion of the energy, particularly in the intense lower-frequency bands. As mentioned previously, the 'silence' frames also have a low residual-to-orginal ratio because the optimal cancellation algorithm is able to cancel a large proportion of the small amount of energy present. These silence frames, which have almost no energy originally, are differentiated from the more energetic solo frames by the second feature, which is related to the original energy.

The post-cancellation residual may be useful for example in sinusoid+noise modeling, and may be useful for classifying the instrument. In ensemble recordings, the comb filter should have removed the most energetic period close to the coarse estimate. This may go some way towards removing lead melodies – a step towards "unmixing" the music. There may also be applications where the coarse period estimate comes from somewhere other than a first-stage autocorrelation. For instance, this kind of cancellation could form part of a score following system that detects the precise timing of anticipated notes by observing when a cancellation filter of the appropriate period effects a large energy reduction.

If the optimal cancellation filter comes out as an approximation to a pure delay – i.e. a conventional comb filter – measuring that delay will give us a very accurate estimate of the dominant pitch in the signal i.e. a high-resolution pitch track. The delay can be sim- ply estimated as a best-fit to the slope of the unwrapped phase of the filter coefficients' Fourier transform. However, we found in many cases these slopes were far from linear.

As discussed in the introduction, our goal was to find a way to accurately and automatically identify solo excerpts within a large music corpus in order to collect training data for solo source models. We believe that the cancellation system is very suitable for this task, and our next step is to apply the system to a large music archive to see what we can find. The ability of the system to detect periodicity without a more detailed model of the particular voice to be found is both a strength and a weakness – it's useful to be able to detect solos for instruments not in the training set, but it means that the returns from the solo detection data mining will themselves need to be clustered and classified to build separate models for distinct instruments.

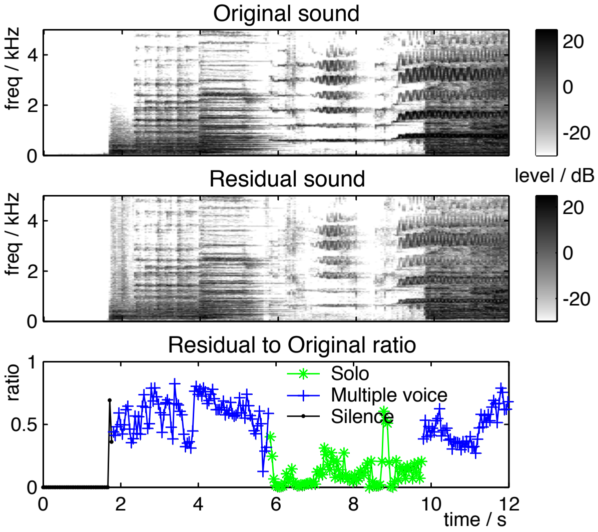


Figure 5: Comparison of a wave file before (top) and after (middle) filtering out the main pitch showing a significant reduction in energy in the area labeled 'single' voice.

5. REFERENCES

[1] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, pp. I–53–56. [Online]. Available: http://www.cs.tut.fi/sgn/arg/klap/source.pdf

[2] A. Berenzweig and D. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, Mohonk, NY, October 2001, pp. 119–122. [Online]. Available: http://www.ee.columbia.edu/~dpwe/pubs/waspaa01-singing.pdf

[3] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.

[4] W. Putnam and J. Smith, "Design of fractional delay filters using convex optimization," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1997.

[5] K. Murphy, "Hidden markov model (HMM) toolbox for matlab," Downloadable software, 2005. [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[6] B. Logan, "Mel frequency cepstral coefficients for music modeling,," in *Proc. International Symposium on Music Information Retrieval*, Plymouth, 2000. [Online]. Available: http://ismir2000.ismir.net/papers/logan_paper.pdf

[7] I. Nabney, *NETLAB: Algorithms for Pattern Recognition*. London: Springer-Verlag London Ltd, 2004. [Online]. Available: http://www.ncrg.aston.ac.uk/netlab/