GEOMETRY IN SOUND: A SPEECH/MUSIC AUDIO CLASSIFIER INSPIRED BY AN IMAGE CLASSIFIER

Norman Casagrande, Douglas Eck and Balázs Kégl
University of Montreal
Department of Computer Science
{casagran,eckdoug,kegl}@iro.umontreal.ca

ABSTRACT

In this paper we adapt a well-known image processing algorithm to the task of predicting whether an audio signal contains speech or music. We derive a frame-level discriminator that is both fast and accurate. Using a simple FFT and no built-in prior knowledge of signal structure we obtain an accuracy of 87% on frames sampled at 20ms intervals. When we smooth the output of the classifier with the output of the previous 40 frames our forecast rate rises to 93% on the Scheirer-Slaney [1] database. To demonstrate the efficiency and effectiveness of the model, we have implemented it as a graphical real-time plugin to the popular Winamp audio player.

1. INTRODUCTION

The ability to automatically discriminate speech from music in an audio signal is useful in domains where a particular type of information is of interest, such as in automatic audio news transcription of a radio broadcast, where non-speech would presumably be discarded. Previous models have employed a mixture of simple features that capture certain temporal and spectral features of the signal [1, 2]. including for example pitch, amplitude, zero crossing rate, cepstral values and line spectral frequencies (LSF). More recently, other approaches have used the posterior probability of a frame being in a particular phoneme class [3], HMMs that integrate posterior probability features based on entropy and "dynamism" [4], and a mixture of Gaussians on small frames [5].

We propose to adapt a successful and robust approach for object detection by Viola and Jones [6] to this task. Our model works by exploiting regular geometric patterns in speech and non-speech audio spectrograms. These regularities are detectable visually, as demonstrated by the ability of certain trained observers to identify speech structure (e.g. vowel formant structure, consonant onsets) and musical structure (e.g. note onsets and harmonic pitch structure) through visual inspection of a spectrogram. We demonstrate in this paper that by exploiting geometric regularities in a two-dimensional representation of sound, we are able to obtain good accuracy results (87%) for 20ms frame categorization with no built-in prior knowledge and at very low computational cost. When smoothing is employed over 40 previous frames (800ms), our accuracy rises to 93%. This compares favorably with other models on the same dataset.

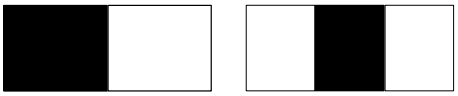

Figure 1. The two Haar-like features used in our additive model.

The use of a vision-inspired model for audio deserves some discussion. We wish to emphasize that despite being motivated by work in vision, this model is well suited for audio signal processing. Though it treats individual 20ms slices of music as having fixed geometry, it places no limitations on the geometry of entire songs. For example, it places no constraints on song length nor does it require random access to the audio signal. In other words, this approach is causal and is able to process audio streams online and in real time.

2. THE ALGORITHM

In order to build a good binary discriminator, it is desirable to find a set of salient *features* that separate the two classes with the largest margin possible. To detect objects in an image, Viola and Jones employed a set of simple *Haar-like* (first proposed by Papageorgiou et al. [7]) rectangles depicted in Figure 1. These features compute and subtract the sum of pixels in the white area from the sum of pixels in the black area. The areas can have different shapes and sizes, and can be placed at different x and y coordinates of the image. A discriminator using a single feature is called a *weak learner* because, used alone, it cannot achieve very good discrimination. However, when these features are combined in an additive model, the resulting classifier can perform very well. In their work on two-dimensional images, Viola and Jones showed that with enough features, it is possible to detect complex objects like faces.

2.1. AdaBoost

To build the additive model, we use the ADABOOST algorithm [8], which is one of the best general purpose learning methods developed in the last decade. It has inspired

several learning theoretical results and, due to its simplicity, flexibility, and excellent performance on real-world data, it has gained popularity among practitioners.

ADABOOST is an *ensemble* (or *meta-learning*) method that constructs a classifier in an iterative fashion. In each iteration, it calls a simple learning algorithm (the *weak learner*) that returns a classifier. The final classification will be decided by a weighted "vote" of the weak classifiers, where each weight is proportional to the correctness of the corresponding weak classifier. If there is no particular a-priori knowledge available on the domain of the learning problem, small decision trees or, in the extreme case, *decision stumps* (decision trees with two leaves) are often used. A decision stump can be defined by three parameters, the index j of the attribute [1] that it cuts, the threshold θ of the cut, and the sign of the decision. Formally,

$$h_{j,\theta+}(\mathbf{x}) = \begin{cases} 1 & \text{if } x^j \geq \theta, \\ -1 & \text{otherwise,} \end{cases} \quad (1)$$

and $h_{j,\theta-}(\mathbf{x}) = -h_{j,\theta+}(\mathbf{x})$. Although decision stumps may seem very simple, when boosted, they yield excellent classifiers in practice. Also, finding the best decision stump using exhaustive search can be done efficiently in $O(nd)$ time, where n is the number of training points, and d is the dimension of the input space (the number of Haar-like features in our case).

For the formal description of ADABOOST, let the training set be $D_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is the observation vector, and y_i is its binary ($+1$ or -1, representing speech or music, respectively) label. The algorithm maintains a weight distribution $\mathbf{w}^t = (w_1^t, \ldots, w_n^t)$ over the data points. The weights are initialized uniformly at the beginning, and are updated in each iteration. The weight distribution remains normalized in each iteration, that is, $\sum_{i=1}^n w_i^t = 1$ for all t. In general, the weight of a point will be proportional to how hard it is to correctly classify. We suppose that we are given a set $\mathcal{H}$ of weak classifiers and a weak learner algorithm that, in each iteration t, returns the weak classifier $h^t \in \mathcal{H}$ that minimizes the weighted error

$$\epsilon^t = \sum_{i=1}^n I_{\{h^t(\mathbf{x}_i) \neq y_i\}} w_i^t, \quad (2)$$

where the indicator function $I_{\{A\}}$ is 1 if its argument A is true and 0 otherwise. The coefficient α^t of h^t is the *confidence* we have in our weak learner. It is set to $\alpha^t = \frac{1}{2} \ln \frac{1-\epsilon^t}{\epsilon^t}$ in each iteration. Since $\epsilon^t < 1/2$ (otherwise we would flip the labels and return $-h^t$) as the algorithm progresses the weight update formulas (for details of the ADABOOST algorithm see [11]) increases the weights of frequently misclassified points, so weak classifiers will concentrate more and more on these "hard" data points. After T iterations [2], the algorithm returns the weighted average $f^T(\cdot) = \sum_{t=1}^T \alpha^t h^t(\cdot)$ of the weak classifiers. The sign of $f^T(\mathbf{x})$ is then used as the final classification of $\mathbf{x}$.

[1] In our case, the jth attribute is the output of the jth filter in the filter bank (see Section 2.2).

[2] T is an appropriately chosen constant that can be set by, for example, cross-validation.

2.2. The features

Our goal is to classify 20ms frames of audio as being either speech or music. We represent each training sample by its spectrogram $S_i = \{S(t, \phi)\}_i$, where $S(t, \phi)$ is the signal intensity at time t and frequency ϕ. We then convolve the image of the spectrogram with Haar-like filters (depicted in Figure1), find the best filter that discriminates the training data, and compute a stump (1) over the output of the filter. Each filter contains two or three rectangular black or white blocks with different sizes and locations. For a black block with its upper left corner placed at (t, ϕ), and with size $w_t \times w_\phi$, we compute the convolution

$$B_{t,\phi,w_t,w_\phi}(S) = \sum_{i=t}^{t+w_t} \sum_{j=\phi}^{\phi+w_\phi} S(i, j).$$

For a white block, we compute the negative convolution $W_{t,\phi,w_t,w_\phi}(S) = -B_{t,\phi,w_t,w_\phi}(S)$. So, for example, a three block white-black-white feature placed at (t, ϕ), and with block size $w_t \times w_\phi$ would output the value

$$W_{t,\phi,w_t,w_\phi}(S) + B_{t,\phi+w_\phi,w_t,w_\phi}(S) + W_{t,\phi+2w_\phi,w_t,w_\phi}(S).$$

The major advantage of these features over more complicated filters usually used in sound-processing that they can be computed at an extremely low cost. The main trick is to pre-compute the so called *integral image* defined as $\Sigma(t, \phi) = \sum_{i=1}^t \sum_{j=1}^\phi S(i, j)$, for each spectrogram in the training sample. Then any convolution B or W can be computed in constant time by using the equation

$$\begin{aligned} B_{t,\phi,w_t,w_\phi}(S) \quad = \quad & \Sigma(t, \phi) + \Sigma(t + w_t, \phi + w_\phi) \\ & -\Sigma(t, \phi + w_\phi) - \Sigma(t + w_t, \phi). \end{aligned}$$

This allows us to evaluate a very large number of candidate features in every boosting iteration. Formally, each Haar-like filter g_j returns a real number $g_j(S_i)$ for each spectrogram S_i, which is the jth attribute x_i^j in the observation vector $\mathbf{x}_i$. Then for each filter, the best decision stump is found. Finally, we select the weak learner h^t which minimizes the weighted training error (2) among all the candidates.

Despite the simplicity of the filters, they can discriminate between speech and music by capturing local dependencies in the spectrogram. For example, the three-block feature depicted in Figure 2 is well-correlated with the speech signal and quasi-independent of the music signal. Figure 3 displays the real-valued output of the filter for all training points, and the threshold of the optimal decision stump. This feature, selected in the first iteration of ADABOOST, has a 30% error rate on the test set.

3. EXPERIMENTAL RESULTS

In the experiments, we used 240 digital audio files of 15 second radio extracts published by Scheirer and Slaney

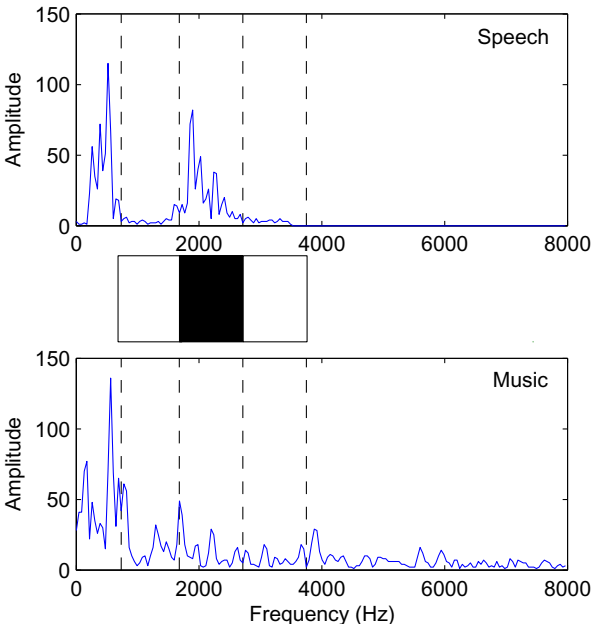

Figure 2. This three-block feature can distinguish speech from music. It is well correlated with the speech signal (its output is 347), and independent of the music signal (its output is -577).

[1][3]. We extracted 11200 20ms frames from this data, and processed them with FFT, RASTA [9], and log-scale FFT. We chose the first because it is the simplest representation of the frequency spectrum, and the other two because of their popularity in speech processing. The FFT represents the biggest frequency space with 256 points, followed by the other two (respectively 26 and 86). The size of this space has a huge impact on the training time since in every iteration, every possible position and block size must be considered. During detection, however, the size of the space does not play any role, due to the integral image representation.

Figure 4 shows the training and test errors using the FFT. The choice for the optimal number of features does not have the same importance as in other machine learning algorithms (e.g., neural networks) because of the intrinsic resistance of AdaBoost to over-fitting: even if the training error tends to zero, the test set error does not increase. It is therefore less important to find a specific stopping point, except for efficiency reasons. We can observe that at a frame level, on a simple FFT we already obtain an error rate of about 13% after 150 iterations, which is far better than the 37% of the best frame-level feature in [1]. Because this representation did not use any information from the past frames, we decided to adapt the classification of a frame to the ones at previous frames with a simple smoothing function. Let $f(\mathbf{x}_\tau)$ be the output of the strong learner, where $\mathbf{x}_\tau$ is a frame at time τ. Then, the

[3] The data was collected at random from the radio by Eric Scheirer during his internship at Interval Research Corporation in the summer of 1996 under the supervision of Malcolm Slaney.

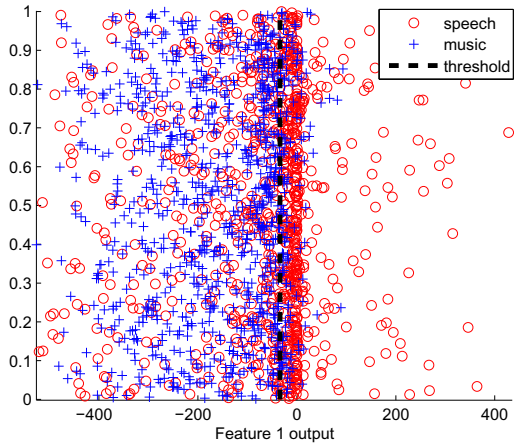

Figure 3. The output of the feature showed in Figure 2 for all training points, and the optimal decision stump's threshold. The data has been randomly distributed on the vertical axis for clarity.

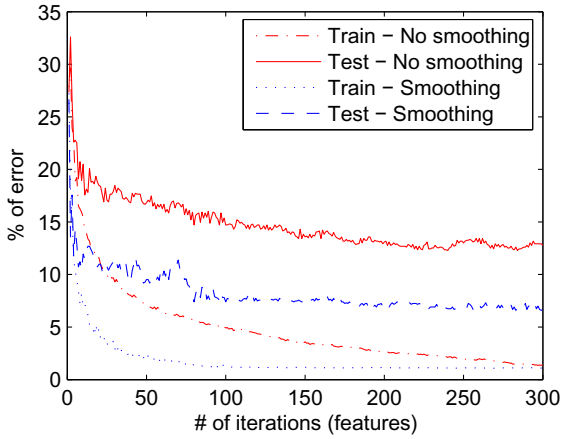

Figure 4. The results on a 20ms FFT frame, without and with smoothing. The benefits of smoothing are clearly seen both in test error and train error.

new output used for classification is

$$g(\mathbf{x}_\tau) = \frac{\sum_{i=max(\tau-nframes+1,0)}^{\tau} a^{\tau-i} f(\mathbf{x}_i)}{\sum_{j=max(\tau-nframes+1,0)}^{\tau} a^{\tau-j}}, \quad (3)$$

where a is a decay parameter between 0 and 1 and where $nframes$ is an integer that corresponds to the number of past frames to consider. In order to find the best values of a and $nframes$, we randomly concatenated the audio (wave) files of the validation set and measured the classification error rates for several values for the decay parameter. We chose this procedure in order to approximately simulate audio streaming from a radio station and get the best values at $a = 0.98$ and $nframes = 40$. With these settings the error reaches a value less than 7% with less than a second of information. The error converges after 150 iterations, but even with a much smaller number of features, such as 75, the error level is below 10%.

	FFT	RASTA	Log-FFT
Without smoothing	12.7%	10.8%	12.6%
With smoothing	6.7%	7.2%	7.4%

Table 1. The error rate using single frame filters.

Surprisingly, RASTA and logarithmic scale FFT representation did not perform as well, even if the results are still below 10%. The training time was, however, a fraction of the training time when the full spectrogram as used, because of their smaller size. Table 1 summarizes the errors on the test set with these representations. Also, RASTA and Log-FFT converged much faster than simple FFT, which can be explained in both cases by the higher quality of the information and the limited dimensionality.

To demonstrate the efficacy of the algorithm, we have implemented a winamp plugin [4] that shows in realtime the discrimination process.

4. CONCLUSIONS

We have showed how a simple generic object recognition algorithm can be used also to perform frame-level classification of audio by exploiting geometric regularities in a fixed-sized two-dimensional representation of frame contents. Because of the strong relationship among frames in time, we can increase the performance of the classifier with a simple smoothing on the output of the frame-level classifier. It is also possible to do training directly on a set of subsequent frames to capture local dependencies in the time domain. However, such an approach would also increase the training time by increasing the size of the search space.

The model is far from being optimized, and further research is necessary to deal well with extremely large training sets. Also it may be helpful to explore the use of different basic features (such as Gaussians or band-passes), and different representations such as wavelets or sine-wave replicas [10].

Finally, while the current model is limited to two-class categorization, we are exploring a multi-class version of ADABOOST [11]. This would allow us to extend our work to more challenging classification problems such as speaker identification singer identification, music instrument identification and music genre classification.

5. ACKNOWLEDGMENTS

We would like to thank Stanislas Lauly for help on data, and Hugo Larochelle for his suggestions. We would like to thank Dan Ellis for helpful comments and for providing the dataset.

[4] Available for free download at `www.iro.umontreal.ca/~casagran/winamp/`.

6. REFERENCES

[1] Eric Scheirer, Malcolm Slaney "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc. of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Volume 2, pp. 1331, 1997.

[2] John Saunders "Real-time discrimination of broadcast speech/music", *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing (Atlanta, GA)*, pp. 993996, May 1996.

[3] Gethin Williams and Daniel P.W. Ellis "Speech/music discrimination based on posterior probability features", *Proc. European Conference on Speech Communication and Technology*, Sept. 1999, pp. 687690.

[4] Jitendra Ajmera, Iain A. McCowan, and Hervé Bourlard "Robust HMM based speech/music segmentation", *Proc. of ICASSP-02*, 2002.

[5] H. Ezzaidi and J. Rouat "Speech, music and songs discrimination in the context of handsets variability", *Proc. of ICSLP 2002, 16-20* September 2002.

[6] Paul A. Viola, Michael J. Jones "Robust Real-time Object Detection", *International Conference on Computer Vision*, pp. 747, 2001.

[7] Constantine P. Papageorgiou, Michael Oren, Tomaso Poggio "A general framework for object detection", *International Conference on Computer Vision*, 1998.

[8] Yoav Freund, Robert E. Schapire "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55(1):119-139, August 1997.

[9] Hynek Hermansky, Nelson Morgan, Arun Bayya, Phil Kohn "RASTA-PLP speech analysis technique", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992. ICASSP-92.

[10] Einat Liebenthal, Jeffrey R. Binder, Rebecca L. Piorkowski and Robert E. Remez "Sinewave speech/nonspeech perception: An fMRI study", *The Journal of the Acoustical Society of America* May 1, 2001 – Volume 109, Issue 5, pp. 2312-2313.

[11] Robert E. Schapire "A Brief Introduction to Boosting", *Proc. of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.