SPECTROGRAPHIC SEAM PATTERNS FOR DISCRIMINATIVE WORD SPOTTING

Shubhranshu Barnwal[1], *Kamal Sahni*[1], *Rita Singh*[2], *Bhiksha Raj*[2]

1. Indian Institute of Technology Kanpur, India
2. Carnegie Mellon University, Pittsburgh, PA, USA
{barnwal, skamal}@iitk.ac.in, {rsingh, bhiksha}@cs.cmu.edu

ABSTRACT

This paper presents a novel method for deriving patterns for classification of speech sounds. In contrast to conventional methods that attempt to capture time-frequency patterns as represented by spectral envelopes or peaks, our method captures patterns of high-energy tracks, or *seams*, of maximum "whiteness" across frequency in spectrograms. Our hypothesis is that these seams could potentially carry relatively invariant signatures of underlying sounds. We present a method to derive feature vectors from seam patterns for discriminative word spotting. We show experimentally that spectrographic seam patterns are indeed distinctive for different spoken words, and are effective for word spotting.

Index Terms— Spectrographic Patterns, Seam Carving, Keyword Spotting, Speech Processing, Hough Transform.

1. INTRODUCTION

Speech recognition systems, and indeed most speech processing applications, attempt to derive, model and classify some basic patterns that characterize units of sound (such as words, syllables or phonemes) that comprise human speech. These patterns are encapsulated in the type of features derived from the speech signal. Different feature representations attempt to capture different aspects of sound-specific patterns in speech. For example, cepstral features represent smoothed spectral envelopes, features based on linear prediction analysis (LPC) attempt to emphasize and capture spectral peaks, with the generally accepted hypothesis that spectral peaks are relatively more characteristic of underlying sounds, while spectral valleys are often affected by noise and carry less classification information. In general, conventional wisdom has been to derive features from spectrographic energy variations, be it peaks or envelopes.

Without negating this conventional wisdom, we analyze a different trend in time-frequency representations of speech sounds: that in continuous high-energy tracks across frequency. Although locality in time is enforced through continuity constraints, the overall tracks can nevertheless span a window of time. We call such a path or trajectory a *seam*. Our core hypothesis is that seams could be least variant across sounds of the same type of sound, and different across different sounds. This paper investigates this hypothesis by analyzing seam patterns across different sounds, specifically spoken words, in controlled experiments.

Seams are not new concepts. In fact, seams of information across images have been effectively used in image processing for resizing or enlarging images, and for changing image content non-linearly without affecting the overall image quality. An extremely successful image processing technique based on seam analysis is called Seam Carving [1]. The technique derives seams through simple dynamic programming across the image plane.

Our method derives from Seam carving, but is more adapted to speech spectrograms. For example, since spectrograms are more tolerant to fine-grained variations in texture, the seams can be smoothed e.g., through a smoothing filter, to yield smoother patterns. Features for each word are then derived using a slight variant of the Hough transform, in a manner similar to edge-detection in images. We then build a discriminative classifier for these features using Support Vector Machines (SVMs) to distinguish between target and non-target words in a simple word-spotting task.

The rest of this paper is arranged as follows: in Section 2 we present the basic spectrographic seam pattern analysis approach. In Section 3 we show how seam patterns can be used to derive features and used for word spotting tasks. In Section 4 we present our experimental results, contrasting them with word spotting based on conventional and alternate spectrographic representations from recent work. In Section 5 we present our conclusions.

2. SPECTROGRAPHIC SEAMS IN SPEECH

For quite a while digital media has had the ability to support dynamic page layouts. By changing the window display size, one changes the layout of the document. However images could only be scaled linearly, often with bizarre results. This changed when Shai Avidan et al. [1] came up with an interesting idea of content aware resizing of images. Their technique was called "Seam carving", and provided a mechanism for resizing images without affecting the aspect of their content. Seam carving functions by establishing a number of seams (paths of least importance) in an image. These can then be removed to reduce image size, or inserted into the image to enlarge it. On an image, depending on the resizing or restructuring goal, seams can be generally horizontal, vertical or differently directed in complex trajectories (e.g. a vertical seam is a path of pixel connected from top to bottom in an image with one pixel in each row).

Audio recordings can conveniently be represented in image form, spectrogram being one of the most common. A path or trajectory on a spectrographic image is defined as a set of connected pixels, and a seam is defined at a trajectory that has the maximum line-integral value. Thus a seam represents a path of maximum spectral "whiteness", or, since whiteness in a spectrum can be related to uncertainty, of maximum uncertainty. However, such paths can be valuable in a pattern recognition sense, since they represent key trends in the underlying image.

2.1. Computing seams

In standard image processing, several types of energy functions can be used for seam computation, *e.g.* gradient magnitude, entropy, visual saliency, eye gaze movement etc. On a spectrogram, each pixel corresponds of a time-frequency bin and represents an energy

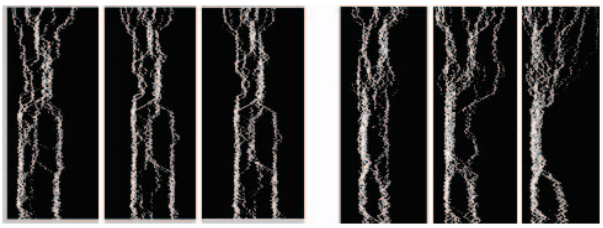

Fig. 1. Low information trajectories are consistent across different instances of a word, but are different across words. 50 seams on two different words - "Elephant" (left 3) and "Tiger" (right 3)

value. Accordingly, we use an energy function for seam computation, where we maximize the energy of each bin along the seam.

Seams are computed using a simple dynamic programming technique. In this, an element of the spectrogram is identified by its position (i, j) in a 2-D matrix of the spectrogram, where i denotes the row index and j denotes the column index. We use just three parameters while computing the seams - element energy denoted by $E[i, j]$, cumulative energy leading up to that element, denoted by $C[i, j]$, and path matrix, denoted by $P[i, j]$, which stores a variable that directs us to previous element's index.

$$C[i,j] = E[i,j] + \max_{k \in [j-2, j+2]} C[i-1,k] \quad (1)$$

At the end of this process, the maximum value of the last row of matrix C indicates the end of the maximum energy containing seam. In the final step, we backtrack from this maximum cumulative energy cell to find the required seam. For ease of description, we will continue to call this seam discovery procedure "seam carving" in the rest of this paper.

The original seam carving algorithm considers only three neighbors for calculating seam paths, as paths had to be connected while removing or inserting seams in an image. Here we opted to consider 5 neighbors as this tends to capture energy variation more explicitly.

2.2. Seam smoothing

Since a spectrogram is much coarser in the information it represents per pixel and in pixel continuity than a standard image of the same size, the seams obtained on a spectrogram are usually jagged. To smooth these out, we use a very basic smoothing filter where a linear penalty is now imposed while computing the seam's path as follows:

$$C[i,j] = E[i,j] + \max_{k \in [j-2, j+2]} Pen[i,k].d.C[i-1,k] \quad (2)$$

where d is deviation distance and $Pen[i, k]$ is a penalty factor that depends on $i - k$.

The penalty imposed not only smoothen the seams, but also forces the seams to not change tracks unnecessarily. Seams are thus rendered more robust in spatial location on the spectrogram. In experiments we noted that penalty values need to vary from keyword to keyword, as some keywords yield smoother seams than others. However, over-penalizing can cause the seam to deviate from its original shape, which is not desirable.

2.3. Fixing the origin

The seams computed are time invariant. There is thus a need for choosing an origin within the seams to enforce uniformity to seams

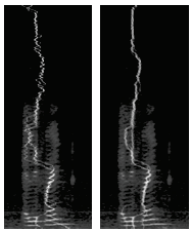

Figure 2. Effect of penalty on seam smoothness. Seams in the right panel are penalized.

obtained from every window. To find this origin we use an approach very similar to finding centre of mass of a given body. X_i are the bin index in time domain, and C_x gives us the location of our origin (in time domain). For the frequency domain or our origin we simply choose the top first row of the spectrogram.

$$C_x = \frac{\sum_i M_i X_i}{\sum_i X_i}; \quad M_i = \begin{cases} 1 \text{ for elements on seams} \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

By using this formula, we ensure that the number of elements that is part of a seam are equally distributed on the two sides. This becomes necessary further, when we use a Hough Transform to extract features for classification from these seams.

3. DERIVING FEATURES FROM SEAMS

Our seam carving algorithm finds a large collection on low-variance seams from spectrograms. The spectrograms are conventional log-magnitude wide-band spectrograms computed using a 25ms analysis window with 10ms frameshift. The individual seams can vary from instance to instance and are not individually useful; rather it is the ensemble that carries information relevant to classification. We use the Hough transform to capture the characteristics of this ensemble.

3.1. The Hough Transform

The Hough transform [2] is a classic feature extraction technique used to identify, or more generally characterize lines and linear structures in images. The basic premise behind the transform is that any line in an image can be represented as a pair of coordinates (r, θ), where r is the length of the normal from the line to the origin and represents the distance of the line from origin, and θ is the orientation of the normal with respect to the X-axis. The relation between the x and y coordinates of any point on the line is given by $xcos\theta + ysin\theta = r$. Viewed alternately, given a point $X = (x, y)$ in an image, the set of *all* lines that pass through X can also be parameterically represented in polar coordinates by the sinusoidal curve

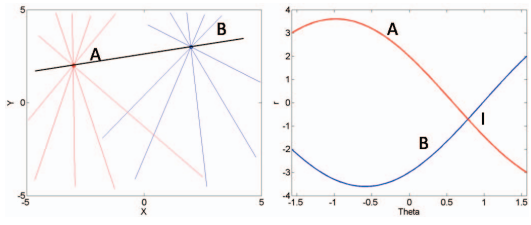

Fig. 3. All lines through the point A in the left image are represented by the red curve in the (r, θ) plane in the right image. All lines through B are represented by the blue curve. The intersection of the two curves I represents the line that goes through both A and B.

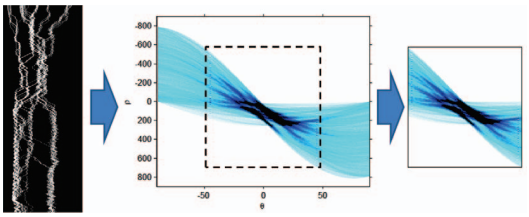

Fig. 4. The collection of seams in the left panel is transformed to the Hough transform in the middle. The central region of this transform, shown to the right, is retained as the feature representing the audio.

$\mathcal{S}_X(r,\theta)$ given by the relation $xcos\theta + ysin\theta = r$. The curve is unique to X. Given two point $X_1 = (x_1, y_1)$ and $X_2 = (x_2, y_2)$, the line that passes through both points is represented by a single point in the $r - \theta$ plane, corresponding to the intersection of the curves $\mathcal{S}_{X_1}$ and $\mathcal{S}_{X_2}$. This is illustrated in Figure 3. The Hough transform builds on this observation. The transform itself is represented in the (r, θ) plane, where $\theta \in [-\pi/2, \pi/2)$ and $r \in \mathcal{R}$. The transform was originally intended to help identify lines and collinear points, *e.g.* edges in images. For every candidate point X_i in the image (where by "candidate" point we refer to points that have, somehow, been identified as possibly belonging to a feature such as an edge), the corresponding curve $\mathcal{S}_{X_i}(r,\theta)$ is plotted on the plane. The contributions of curves derived from individual points is additive. Thus, if the curves corresponding to K candidate points $\mathcal{S}_{X_i}(r,\theta), i = 1 \ldots K$ intersect at some point (R, Θ) in the (r, θ) plane, the transform $\mathcal{H}(R, \Theta) = K$. Also, conversely if $\mathcal{H}(R, \Theta) = K$ at some (R, Θ), then it represents K candidate points in the image that are collinear. In practice, the Hough transform quantizes the (r, θ) space. The output of the transform is a matrix, whose axes represent quantized r and θ.

3.2. Characterizing seams through the Hough transform

A Hough transform that is computed over a large number of candidate points, when viewed as an image itself, will exhibit several regions of high intensity and others of low intensity. The high intensity regions represent groups of collinear points, where the higher the intensity, the larger the number of points in the group. Regions of low intensity represent small sets of points that are mutually collinear, but not aligned to other points. The complete transform therefore encodes the overall arrangement of the candidate points.

To characterize the collection of seams obtained from a spectrogram, we therefore compute a Hough transform from all points on all detected seams. This is illustrated in Figure 4. We make two adjustments. First, the origin of the transform is taken to lie at frequency = 0, and at the time location that represents the horizontal center of mass of all points on all seams (which is computed as explained above). Second, the majority of the regions in the transform are relatively low in intensity and contain little information about the overall seam pattern. We therefore only retain the central region of the transform as also illustrated in Figure 4. The retained central portion of the Hough transform is a matrix of numbers, which is unravelled into a "seam-hough" feature vector.

4. CLASSIFICATION

Classification is performed using a Support Vector Machine (SVM) [3]. To train the SVM we compute seam-hough features from several segmented out instances of the word as positive exemplars. Negative training exemplars are seam-hough features obtained from randomly drawn segments of speech from recordings that do not contain the target word. The width of negative examples is distributed similarly to that of positive examples.

For the ROC curve of Figure 5 we merely classified segmented instances of words using the SVM. For the larger task of word spotting we computed features from and classified a sliding window, whose width was set to be 1.2 times the mean width of the word (in the training data), using the SVM. The hopsize between adjacent windows was 20% of the size of the window. When adjacent windows classified the underlying audio as the word, we merged the segments into a single unit detection. If any such hypothesized segment exceeded twice the average length of the word (as obtained from the training data), we divided it into multiple segments, each of length equal to twice the average length of the word, and each resulting segment was called a separate detection. Since, in reality, spoken words may be more than twice their average length, this can result in a pessimistic calculation of detection performance. In all cases, in order to compute the entire ROC the distance of the instances from the boundary was compared to a varying threshold.

5. EXPERIMENTS

5.1. Training, testing and comparators setup

We ran experiments on the TIMIT speech corpus (available from LDC) to evaluate our technique. We extracted 462 examples of the keywords "greasy", "dark", "wash", and "oily" from this corpus. We then divided this into two sets, one to be used for training and one for testing. Each of these two subsets was further subdivided to positive sets and negative sets, the former included only the target keyword, while the latter included utterances which did not contain the target keyword. The number of elements in the positive set were equal to number of elements in the negative set while training of each target keyword's SVM model. For the SVM, we used the linear kernel.

We note that our features are rather unconventional, and are best categorized as "alternative" spectrographic features. Other alternate spectrographic features have recently been explored by researchers and found effective for simple word spotting tasks. E.g., in [4], Ezzat explored a mechanism based on detection of patches of spectro-temporal features for keyword spotting. Ezzat's method also derives non-standard word-level features from spectrograms, and has repeatedly demonstrated results comparable or superior to HMMs under low-training data situations in particular.

The benefit (and reason) for choosing the particular experimental setup that we did is that it is identical to the one used by Ezzat et. al. in their experiments. Thus, it enables us to compare our results on a one-to-one basis with theirs, not only with the performance they obtain with their method, but also with their *baseline*, which, presumably, was optimized appropriately. Thus, in the results reported below, the comparative results reported, namely the HMM-based baseline and the performance obtained by Ezzat *et. al.* are directly drawn from their paper, enabling us to report what we believe to be fair comparisons.

In a first experiment we evaluated the effect of the *number* of seams derived from the spectrogram on classification performance. Figure 5 shows the ROCs obtained for the word "dark" with different numbers of seams. Here we classified segmented-out positive and negative instances of the word. We note that increasing the number of seams generally improves classification. In the remaining experiments we used 25 seams to compute seam-hough features.

Figures 6 and 7 shows the ROC curves obtained for each of

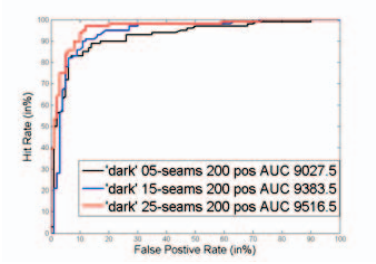

Fig. 5. The ROCs obtained with different numbers of seams.

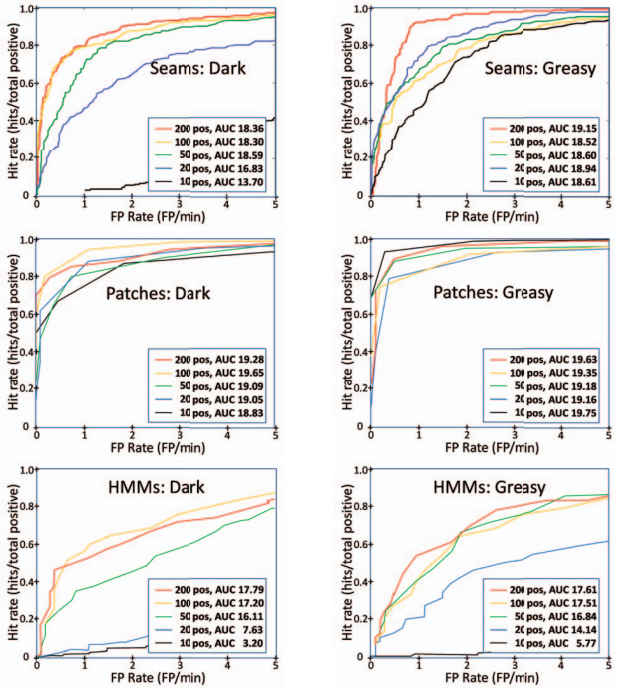

Fig. 6. Results on "Dark" and "Greasy".

the four words "greasy", "dark", "wash" and "oily", for classifiers trained with different numbers of training instances. In these plots the y axis shows the percent of target words that were correctly detected, while the x axis shows the number of *false* detections performed per unit time. Also shown are the results reported by Ezzat et. al. in [4] for exactly equivalent experiments using their patch-based spotter, and an HMM-MFCC based comparator.

6. OBSERVATIONS, CONCLUSIONS AND FUTURE WORK

We note that the performance obtained with seams-based features is nearly as good as that reported by Ezzat *et. al.* using spectral patches, and generally better than the performance obtained with an HMM-MFCC classifier. What is significant is not that we do not *outperform* Ezzat *et. al.* – indeed we believe that the HMM-based approach suffers primarily from the fact that it is not discriminatively optimized for word spotting. It is that although the seams only detail the *location* of high-energy tracks in the spectrogram, and do not represent the spectral information itself, we nevertheless get comparable performance to that obtained with MFCCs or spectrographic

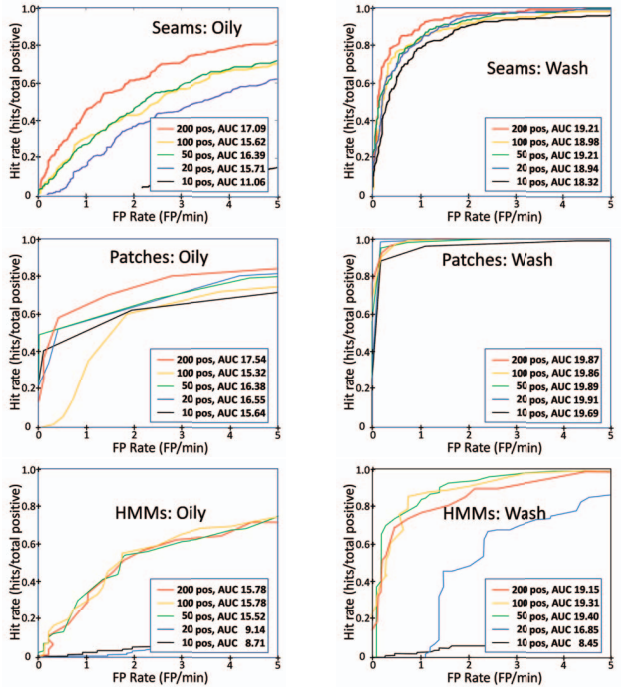

Fig. 7. Results on "Oily" and "Wash".

patches which characterize spectral shape. The two features capture different kinds of information. It is reasonable to assume that by combining the two one can obtain better performance still.

Another observation is that the seam features could be expected to be most useful for words containing consonants, stops and other constituents that are likely to impart distinctive vertical seams to the spectrogram. Words such as "oily" which are entirely sonorant do not have distinctive vertical seams. Yet, by forcing the algorithm to discover seams we do derive information that is sufficient to recognize "oily" no worse than the other methods. We have not investigated all possibilities: horizontal seams, not tried in this paper, may have additional information, that could be valuable for classification.

We are conducting detailed investigations of the seam features obtained from different types of sounds, *i.e.*, vowels, voiced and unvoiced stop consonants and fricatives, and the effect of context, coarticulation etc. In addition, we are investigating the relationship between spectrographic seams and various perceptual cues. We also intend to investigate other ways of abstracting seam based representations besides the Hough transforms, as well as techniques for harnessing seam structure within continuous speech recognition tasks.

7. REFERENCES

[1] Shai Avidan and Ariel Shamir, "Seam carving for content aware resizing," in *SIGGRAPH*, 2007.

[2] R. O. Duda. and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," in *Comm. ACM, Vol. 15*, January 1972, pp. 11–15.

[3] Nello Cristianini and Bernhard Scholkopf, "Support vector machines and kernel methods: the new generation of learning machines," in *AI Magazine, Vol. 23, No. 3*, 2002, pp. 31–41.

[4] Tony Ezzat and Tomaso Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *SAPA*, 2008.