Shift-Invariant Features for Speech Activity Detection in Adverse Radio-Frequency Channel Conditions

Mohamed Kamal Omar, Sriram Ganapathy

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
mkomar,sganapathy@us.ibm.com

Abstract

This work presents a novel approach to speech activity detection for highly degraded radio-frequency channel conditions. In this approach, the audio stream is segmented into short homogeneous segments. Each segment is represented by shift-invariant features. These features provide a coarse histogram-based description of the high-energy trajectories in the time-frequency domain. They are less sensitive to frequency shifting compared to traditional filterbank-based features like Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients. We evaluate our approach on the speech activity detection task of the Robust Automatic Transcription of Speech (RATS) program. Our experiments show improvements up to 29% relative in the performance in terms of total error on four radio-frequency channels used in RATS compared to the PLP-based baseline system.

Index Terms: speech activity detection, segmental modeling, invariant features

1. Introduction

Speech activity detection (SAD) is an important preprocessing step for audio analytics. SAD in highly degraded conditions is still a challenge for many applications [1, 2, 3, 4]. The challenge is even harder when the training data used to estimate the parameters of the SAD models is mismatched with the testing condition which is the focus of this work.

We address this problem in the context of the Robust Automatic Transcription of Speech (RATS) program. The program targets audio analytics in extremely noisy and highly distorted radio-frequency channels [5]. One of the motivations of this work is the observation that human annotators are very successful in segmenting RATS audio to speech and non-speech regions using the spectrogram. Spectrogram reading in this case involves interpreting the acoustic patterns in the spectrogram to determine simultaneously the boundaries between different segments and their labels as speech or non-speech. The basic idea here is that pitch and formant trajectories exhibit patterns which can be differentiated from the mostly random or even harmonic structures in the non-speech regions. This motivates developing a segment-based approach as an alternative to the conventional frame-based approaches [1, 2, 6].

In this work, the audio stream is segmented into short segments by detecting the change points, these segments are represented using shift-invariant features inspired by the work on shape detection in [7]. By employing a representation that is shift-invariant in time, we can compensate for many inaccuracies in generating the segments. By employing a representation that is shift-invariant in frequency, the representation may be robust to some types of carrier mismatch between the transmitter and the receiver and tonal variations in the RATS data [5]. Figure 1 shows the spectrogram of a speech segment before and after transmission through one of the RATS channels. This robustness, as we will discuss later, comes at the expense of losing some of the detailed filterbank-based frame-based information employed in standard features like MFCC and PLP.

In the next section, we describe the task, the data, and the baseline system. In Section 3, the shift-invariant segment-based approach is introduced. The experiments performed to evaluate the different techniques are described in Section 4. Finally, Section 5 contains a discussion of the results.

2. Setup

In this section, we provide a brief description of the data used in training and testing and the baseline system.

2.1. Data

This work is part of our efforts on the SAD task of the RATS program. Both the training and the testing data in the RATS program consist of audio recordings from a variety of existing data sources, transmitted through LDC's multi radio-link channel collection system with fixed transmitter and receiver settings [5]. These recordings comprise simultaneous captures from different transmitter/receiver combinations. These combinations represent different kinds of modulation, carrier channel bandwidth, receiver intermediate frequency (IF) bandwidth, and range of carrier frequencies mostly in the high frequency (HF) and the ultra high frequency (UHF) ranges. Sources of distortion like sideband mistuning, multipath interference, and tonal interference are very common. Table 1 shows the configuration of the four RATS channels presented in this work [5]. The total amount of training data used to train each of our systems is approximately 800 hours. The systems are evaluated on two test sets: RATS dev1 and dev2 data sets. The dev1 test set consists of approximately 1.5 hours per channel. The dev2 evaluation data is approximately 2 hours per channel. The metrics used in the RATS evaluation are both the false alarm rate and the missing probability. The results are presented here in terms of the two metrics in addition to their average, the total error.

2.2. The Baseline system

The baseline system is a two-pass system which uses the initial pass to adapt the parameters that are used in the final pass to generate the final speech/non-speech segmentation [8], while the proposed system is a single-pass system which processes the input audio only once to generate the final speech/non-speech segmentation.

In the baseline system, two speech and non-speech Gaus-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

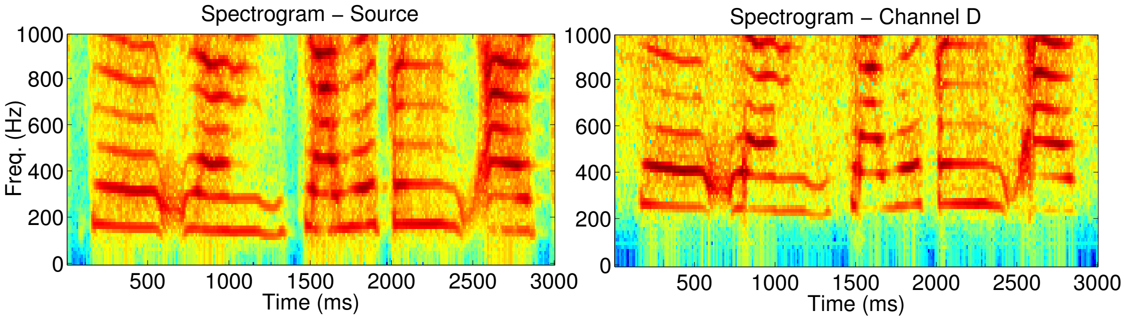

Figure 1: *An example of the frequency shifting effect of the RATS channel D.*

Table 1: The configurations of the four RATS channels used in our experiments

Channel	Transmitter	Receiver	Receiver IF Bandwidth	Transmission Band	Modulation	Channel Bandwidth
B	Icom ICF21GM	AOR AR5001D	6KHz	UHF	NFM	6.25KHz
D	Galaxy DX2547	Icom IC-R75	9KHz	HF	SSB	10KHz
E	Icom IC-F70D	Icom ICR8500	6KHz	VHF	NFM	11.25KHz
H	Magnum 1012HT	TenTec RX340	15KHz	HF	NFM	10KHz

sian mixture models (GMMs) are used for initial segmentation and speech/non-speech labeling of the input recording [9]. The speech and non-speech segments with high confidence are used to MAP adapt the means of the Gaussian components of the corresponding GMMs [10]. These adapted models are used in a final pass to provide the final speech/non-speech segmentation of the input recording [8].

3. The Shift-Invariant Segment-Based Approach

In this section, we discuss the shift-invariant segment-based approach. First, the input audio is segmented into short segments and then a representation of each of these segments is estimated. An MLP model is trained using the segments of the training data. In the following, we provide the details of each of these steps.

3.1. Segmentation of the input audio

The main goal of this step is to create homogeneous segments that contain only speech or non-speech regions but not both. The algorithm uses several heuristics to detect a change point. One of these heuristics is

$$TD[n,f] \quad = \quad \min_{\rho \in N(f)} \left| x[n,f] - x[n+1,\rho] \right|, \qquad (1)$$

where $N(f)$ is a small window in the frequency domain around f, and $x[n,f]$ is the value of the logarithm of the magnitude spectrum at frame n and frequency index f. Three other measures are used in our algorithm to detect segment boundaries. The first estimates the minimum difference at the same frame across a small window in the frequency domain around f,

$$SD[n,f] \quad = \quad \min_{\rho \in N(f)} \left| x[n,f] - x[n,\rho] \right|. \qquad (2)$$

The second estimates the minimum difference with the previous frame across a small window in the frequency domain around f,

$$PD[n,f] \quad = \quad \min_{\rho \in N(f)} \left| x[n,f] - x[n-1,\rho] \right|. \qquad (3)$$

The third measure uses a window that spans six frames around the current frame and is given by

$$\begin{aligned} LT[n,f] \quad &= \quad \sum_{m=n-3}^{n-1} x[m, f+(m-n)(f-v)] \\ &\quad - \sum_{m=n+1}^{n+3} x[m, f+(m-n)(f-v)], \quad (4) \end{aligned}$$

where

$$v \quad = \quad \operatorname*{argmin}_{\rho \in N(f)} \left| x[n,f] - x[n+1,\rho] \right|.$$

The heuristic measurements described in Equations 1, 2, 3, and 4 are used to generate three change indicators at each frame as follows

$$TDI[n] \quad = \quad \#(TD[n,f] > Th_1), \qquad (5)$$

where $\#(C[n,f])$ is a function that returns the number of points in the time-frequency domain at which the condition $C[n,f]$ is true, and Th_1 is a threshold selected using a held-out set. The second change indicator is

$$LTI[n] \quad = \quad \#(LT[n,f] > Th_2), \qquad (6)$$

where Th_2 is a threshold selected using a held-out set. The third change indicator is

$$SPI[n] \quad = \quad \#(SD[n,f] + PD[n,f] > 2TD[n,f]). (7)$$

The three values in Equations 5, 6, and 7 are added together and the sum is compared to a threshold, Th_3, to determine if there is a change point at frame n or not.

The threshold values are estimated on a held-out set of 50 10-minute files from the RATS SAD training data to minimize a weighted sum of the missing probability and false alarm. To achieve the goal of creating homogeneous segments and since segments can be merged after detecting that they belong to the same class, we selected an operating point at which the false alarms are twice the missing errors. To avoid extremely large number of segments, we have a limit on the minimum duration of the segment equals 0.2 second. The values used for Th_1, Th_2, and Th_3 in our experiments are 3.9, 8.2, and 80 respectively. These are kept constant across all experiments on different channels.

3.2. Estimation of the segmental representation

Each segment is represented in our algorithm by 100 points in the time-frequency domain. We experimented with selecting these points using several criteria. For example, we tested using an edge detector algorithm [11] to select these points as described in [7]. However, our best results were achieved by selecting the points corresponding to the highest log magnitude spectrum values in the segment. Each point p_i is represented by a coarse histogram of the relative coordinates of the remaining selected points as in [7]

$$h_i(k) \quad = \quad \#\{q \neq p_i \colon (q - p_i) \in bin(k)\}\,. \qquad (8)$$

We experimented with bins that are uniform in the log-polar space, making the representation more sensitive to the close sample points than to points farther away as in [7]. However, the best results were achieved by using independent bins for each dimension with the bins uniformly distributed.

3.3. The MLP-based segmental classifier

We use an MLP with one hidden layer as a classifier to determine the label of each segment. An MLP-based classifier was used instead of the GMM used in the baseline system because the features in this case have high correlations which violate the diagonal covariance assumption. Also a full-covariance model is not efficient because of the high dimensionality of the features. For each segment, the input to the MLP is the concatenation of the representation of the selected 100 points. The order of these points in the representation has a major effect on the performance. We explored various approaches. This includes ordering the points based on time and then frequency indices for points that are from the same frame, and ordering the points based on frequency index and then based on time for points that have the same frequency index. Also we tried reordering the points to reduce a measure of the variance across adjacent points. However, the best performance was achieved by ordering the points based on frequency and then based on time for points that have the same frequency index. The estimation of the MLP parameters is performed using the Quicknet toolkit [12] with the minimum cross entropy objective function.

4. Experiments

In this section, we describe the implementation of the different features and the experiments performed to evaluate them.

4.1. Implementation

In both the baseline and shift-invariant segment-based systems, the input audio is down sampled from 16 KHz to 8 KHz and then windowed to frames of 32 ms duration with a shift of 10 ms. For the baseline system, thirteen PLP coefficients are calculated for each frame. Cepstral mean normalization is then applied per utterance. The PLP coefficients of 9-frames around the current frame are spliced together and then projected to a 40-dimensional vector using linear discriminant analysis (LDA). Each GMM consists of 4042 diagonal-covariance Gaussian components. The parameters of the two models are estimated from the training data using maximum likelihood estimation. For the shift-invariant segment-based system, the fast Fourier transform (FFT) of size 256 is estimated. Only values in the range of 125 to 3800 Hz are used, as almost all of the RATS channels did not have significant energy outside this range. This gives 118 values for each frame. After segmenting the utterance to short segments as discussed before, the 100 points in the time-frequency domain with the highest log magnitude spectrum values are selected to represent the segment. We experimented with many configurations for representing these points. We report here the two most successful representations:

Table 2: Comparing the shift invariant and the scale and shift invariant representations of the segment on the RATS dev1 test set

	Shift-Invariant			Scale-and-Shift-Invariant		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	4.1	8.7	6.4	4.0	8.4	6.2
D	3.4	5.7	4.5	3.7	5.9	4.8
E	3.4	9.9	6.6	4.0	9.8	6.9
H	5.5	5.1	5.3	5.9	5.6	5.7

1. The histogram representation: Each point is represented by the histogram of the relative values of the remaining points, as in Equation 8, along the dimensions of time, frequency, and log magnitude spectrum. We used 4 bins along each dimension. This gives 1200-dimensional representation of each segment.
2. The mixed representation: Each point is represented by the histogram of the relative values of the remaining points along the dimensions of time and frequency, in addition to the zero-mean normalized log magnitude spectrum, the estimate of the time derivative, $TD[n, f]$, as in Equation 1, and the average of $TD[n, f]$ at frame n for all frequency indices f with log magnitude spectrum values higher than the average value across the segment, $\frac{1}{|F|}\sum_{f\in F} TD[n, f]$, where $F = \{f : x[n, f] > x_{avg}\}$ and x_{avg} is the average log magnitude spectrum across the segment.

In both cases, the feature vector is used as an input to train an MLP with one hidden layer. We experimented with three sizes of the hidden layer: 500, 800, and 1200. The size of the output layer is two nodes corresponding to the speech and the non-speech labels. The posterior of the speech class generated by the MLP is compared to a threshold to decide on which label will be assigned to the segment. The threshold is estimated on a held-out set of 50 10-minute files from the SAD RATS training data to get an operating point close to the equal error rate (EER) region. The value of the threshold in our experiments is 0.63.

4.2. Results

Our focus in this work is on evaluating the performance of the two systems on data corresponding to a channel not used in the training data. However, we first evaluate the segment-based system with the testing channel included in training. In the first set of experiments, we tested the effect of making the segment representation scale invariant by dividing the values for each point by the range of possible values over all the selected points of the segment. We used the mixed representation with a hidden layer of 800 nodes in these experiments. As shown in Table 2, the results are mixed with small improvement on channel B from scale invariance and slight degradation on channels D, E, and H. In all the experiments reported in the rest of the paper, we choose to use the shift invariant representation of the segment.

In Table 3, the performance of the shift-invariant segment-based system with the mixed representation is compared for different sizes of the MLP hidden layer. The results show that across most channels, significant gains are obtained by using

Table 3: Comparing the proposed system with different sizes of the MLP hidden layer on the RATS dev1 test set

	500			800			1200		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	4.3	8.8	6.5	4.1	8.7	6.4	4.1	8.6	6.3
D	5.2	5.6	5.4	3.4	5.7	4.5	3.4	5.7	4.5
E	4.9	9.8	7.3	3.4	9.9	6.6	3.5	9.2	6.3
H	5.7	5.9	5.8	5.5	5.1	5.3	5.8	5.3	5.5

Table 4: Comparing the histogram and the mixed representations of the segment on the RATS dev1 test set

	Histogram Representation			Mixed Representation		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	3.9	7.9	5.9	4.1	8.7	6.4
D	3.9	6.2	5.0	3.4	5.7	4.5
E	4.2	9.8	7.0	3.4	9.9	6.6
H	6.2	6.3	6.2	5.5	5.1	5.3

Table 5: Comparing the proposed segment-based and the two-pass baseline system on the RATS dev1 test set with matched training

	Proposed			Baseline		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	4.1	8.7	6.4	5.3	5.6	5.4
D	3.4	5.7	4.5	2.0	6.6	4.3
E	3.4	9.9	6.6	2.9	9.0	5.9
H	5.5	5.1	5.3	3.0	5.4	4.2

800 hidden nodes instead of 500. Comparing the results of using 800 hidden nodes and 1200 hidden nodes shows that the results are not consistent across channels. We report results using the 800 hidden node system in the rest of the paper.

In Table 4, we compare the histogram and the mixed representations. As shown in Table 4, using explicit magnitude-spectrum-related values as in the mixed configuration improves the performance on channels D, E, and H, but slightly degrades the performance on channel B. All the results reported in the rest of the paper are based on the mixed representation.

The performance of the single-pass segment-based system is compared to the two-pass frame-based MAP-adaption baseline system in Table 5. The results indicate that when the system is trained on the testing channel in addition to the other channels, the baseline two-pass MAP-adaption system outperforms the proposed shift-invariant segment-based system by 4% to 26% relative. However, this improvement is gained at the expense of approximately three times slower detection as the baseline system has larger models and uses two-pass decoding to generate the final segmentation.

In Table 6, the comparison of the results of the proposed shift-invariant segment-based system and the two-pass frame-based MAP-adaptation baseline system on the dev1 test set show that the proposed system consistently outperforms the two-pass baseline system across the four channels, when the test channel is not included in the training data. The gain in the performance from using the proposed shift-invariant segment-based system compared to the baseline system ranges between 9% to 29% relative in terms of total error. This may indicate that the features used in the proposed system are more robust to frequency shifting and limited tonal variations in these channels than the traditional PLP coefficients in the baseline system.

Table 6: Comparing the proposed segment-based and the two-pass baseline system on the RATS dev1 test set with mismatched training

	Proposed			Baseline		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	5.5	10.6	8.0	9.2	10.9	10.0
D	5.2	5.2	5.2	4.9	9.7	7.3
E	9.2	8.9	9.0	9.1	10.7	9.9
H	5.5	7.9	6.7	5.9	9.6	7.7

Table 7: Comparing the proposed segment-based and the two-pass baseline system on the RATS dev2 test set with mismatched training

	Proposed			Baseline		
Channel	Miss (%)	FA (%)	Total (%)	Miss (%)	FA (%)	Total (%)
B	7.4	9.6	8.5	8.7	12.8	10.7
D	5.7	6.1	5.9	6.5	9.8	8.1
E	6.1	5.3	5.7	5.5	7.3	6.4
H	4.9	7.7	6.3	5.1	9.5	7.3

Finally, we compare the results of the proposed shift-invariant segment-based system and the two-pass baseline system on the dev2 test set across the four channels when the test channel is not included in the training data in Table 7. The gain in the performance from using the proposed shift-invariant segment-based system compared to the two-pass baseline system ranges between 11% to 28% relative in terms of total error which is consistent with the results on the dev1 test set.

5. Discussion

In this paper, we examined a novel approach to automatic detection of speech in noisy radio channels. This approach involves segmenting the audio into short segments and then representing each segment with shift-invariant features. Consistent improvements across the four channels on both the RATS dev1 and dev2 test sets are achieved compared to the baseline system, when the data of the testing channel is not used in training.

In the proposed system, filters with fixed center frequency and bandwidth which are used in traditional speech processing frontends like PLP and MFCC are avoided. This makes the features less sensitive to frequency shifting and tonal variations typically encountered in radio-frequency channels. This may explain the better generalization in the mismatched condition at the expense of a degradation in the performance when the testing channel is used in training in the matched condition setup.

6. Acknowledgments

This work is supported in part by contract no. D11PC20192 DOI/NBC under the RATS program. The views, opinions, findings and recommendations contained in this article are those of the authors and should not be interpreted as representing the views or policies, either expressed or implied, of the DOI/NBC.

7. References

[1] P. K. Ghosh, A. Tsiartas, S. Narayanan, "Robust voice activity detection using long-term signal variability," in *IEEE Transactions On Audio, Speech, and Language Processing*, vol. 19, no. 3, March 2011.

[2] H. Ghaemmaghami, B. Baker, R. Vogt, S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. of InterSpeech*, pp. 1372–1375, 2010.

[3] A. Temko, D. Macho, C. Nadeu, "Enhanced SVM training for robust speech activity detection," in *Proc. of ICASSP*, vol. 4, pp. 1025–1028, April 2007.

[4] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filtering," in *Proceedings of Interspeech*, pp. 2933–2936, August 2007.

[5] K. Walker, S. Strassel, "The RATS Radio Traffic Collection System," in *Proc. of Speaker Odyssey*, June 2012.

[6] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a Speech Activity Detection System for the DARPA RATS Program," in *Proc. of Interspeech*, September 2012.

[7] S. Belongie, J. Malik, J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," in *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 24, no. 24, April 2002.

[8] M. K. Omar, "Speech Activity Detection for Noisy Data Using Adaptation Techniques," in *Proc. of Interspeech*, September 2012.

[9] M. K. Omar, U. Chaudhari, G. Ramaswamy, "Blind change detection for audio segmentation," in *Proc. of ICASSP*, pp. 501–504, April 2005.

[10] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[11] H. Moon, R. Chellappa, and A. Rosenfeld, "Optimal Edge-Based Shape Detection," *IEEE Transactions on Image Processing*, vol. 11, no. 11, pp. 1209–1226, 2002.

[12] *http://www.icsi.berkeley.edu/Speech/qn.html*.