

ROBUST FEATURE CLUSTERING FOR UNSUPERVISED SPEECH ACTIVITY DETECTION

Harishchandra Dubey, Abhijeet Sangwan, John H. L. Hansen⁺

Robust Speech Technologies Lab, Center for Robust Speech Systems
The University of Texas at Dallas, Richardson, TX 75080, USA

{Harishchandra.Dubey, Abhijeet.Sangwan, John.Hansen}@utdallas.edu

ABSTRACT

In certain applications such as zero-resource speech processing or very-low resource speech-language systems, it might not be feasible to collect speech activity detection (SAD) annotations. However, the state-of-the-art supervised SAD techniques based on neural networks or other machine learning methods require annotated training data matched to the target domain. This paper establish a clustering approach for fully unsupervised SAD useful for cases where SAD annotations are not available. The proposed approach leverages Hartigan dip test in a recursive strategy for segmenting the feature space into prominent modes. Statistical dip is invariant to distortions that lends robustness to the proposed method. We evaluate the method on NIST OpenSAD 2015 and NIST OpenSAT 2017 public safety communications data. The results showed the superiority of proposed approach over the two-component GMM baseline.

Index Terms: Clustering, Hartigan dip test, NIST OpenSAD, NIST OpenSAT, speech activity detection, zero-resource speech processing, unsupervised learning.

1. INTRODUCTION

Speech activity detection (SAD) is an essential front-end in most speech systems such as automatic speech recognition, speaker verification *etc* [1]. SAD methods are broadly considered into two categories: (1) supervised and (2) unsupervised. While supervised approaches are trained on massive amount of annotated data, unsupervised techniques do not require labeled data [2]. Supervised techniques tend to perform poorly on mis-matched train and test conditions. Gaussian mixture models (GMMs) have been extensively used for supervised, semi-supervised and unsupervised SAD [1, 3, 4]. Robust SAD over degraded channels have been of interest for several years [5, 6, 7, 8, 9, 10, 11]. SAD methods are varied, from energy-based [2] to deep neural networks (DNN) [7]. The DARPA RATS program supported the SAD research in

multiple phases that led to the development of advanced approaches [12, 13, 14, 15, 16, 17]. Recent work in [1] summarized the SAD developments in context of semi-supervised and unsupervised techniques. Specifically, it introduced the idea of semi-supervised learning in conventional expectation-maximization (EM) algorithm for semi-supervised GMM for speech activity detection.

2. PROPOSED METHOD

2.1. Feature Extraction

The handcrafted five-dimensional features for Combo-SAD approach were introduced in [3]. Authors performed mean and variance normalization on each feature dimension. The normalized features were later processed with principal component analysis (PCA) for extracting the first principal component that was named *Combo* feature. The Combo features were later employed to consider a two-component GMM for unsupervised SAD [3]. We used the two-component GMM as baseline decision backend for comparison with proposed Dip-based unsupervised backend in this study.

2.2. Hartigan dip test

The dip test [19] is a statistical test for hypothesizing the modality of a distribution. It is based on the geometrical shape of the feature distribution. The dip test tries to fit a piecewise linear function, that is convex then a concave, to the cumulative distribution. The unimodality is decided based on the *goodness of this piecewise linear fit* [18]. We leveraged recursions based on dip test for clustering feature space into speech and non-speech classes. This paper is motivated by the recent success in applying Hartigan test for clustering extremely noisy data from other domains [20]. Application to speech processing, particularly speech activity detection is a novel contribution of this paper. By comparing the dip statistics with that of a suitable reference unimodal distribution (i.e., null distribution), a p-value is set for the null hypothesis. Using the significance level, $\alpha = 0.05$, we may reject or favor the null hypothesis (unimodality) against the alternative hypothesis (multi-modality). In this way, the dip test quantifies the empirical cumulative distribution's departure from

⁺This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

Algorithm 1 *computeDip*

Input: speech features were sorted in ascending order i.e., $\mathbf{o}=[o_1, o_2, \dots, o_N]$ where $o_1 \leq o_2 \leq \dots \leq o_N$.

Output: primary modal interval $[o_L, o_U]$, DIP and p-value, p .

Step 1: Initialize, lower point $o_L = o_1$, upper point $o_U = o_N$ and $D = 0$.

Step 2: Compute greatest convex minorant G and least concave majorant H of empirical distribution F of features in interval $[o_L, o_U]$ [18]. Let the points of contact with F are respectively, g_1, g_2, \dots, g_k (for G) and h_1, h_2, \dots, h_m (for H).

Step 3: Let $d = \max |G(g_i) - H(g_i)| > \max |G(h_j) - H(h_j)|$ and the maximum occurs at $h_j \leq g_i \leq h_{j+1}$. Then, define $o_L^0 = g_i, o_U^0 = h_{j+1}$.

Step 4: Let $d = \max |G(h_j) - H(h_j)| \geq \max |G(g_i) - H(g_i)|$ and the maximum occurs at $g_i \leq h_j \leq g_{i+1}$. Then, define $o_L^0 = g_i, o_U^0 = h_j$.

Step 5: If $d \leq D$, Stop and set $DIP = \frac{D}{2}$.

Step 6: If $d > D$, set $D = \max \{$

$$\sup_{o_L \leq o \leq o_L^0} |G(o) - F(o)|, \sup_{o_U^0 \leq o \leq o_U} |H(o) - F(o)|$$

$\}$, where sup is the supremum (supremum is the smallest number that is greater than or equal to every number in the set).

Step 7: Set $o_L = o_L^0, o_U = o_U^0$. Go to Step 2.

unimodality. Importantly, the dip test (see Algorithm 1 *computeDip*) communicates the modal interval $[o_L, o_U]$, the p-value and the DIP. It is important to note that the proposed clustering approach works on all frames of a single utterance thus it a utterance-level approach. The speech feature vector, **feats** are sorted in increasing order. We still store the original feature vector in memory for preserving the temporal order (time information) of the frames. Let the sorted features (observations) be $\mathbf{o} = o_1, o_2, \dots, o_N$ with $o_1 \leq o_2 \leq \dots \leq o_N$ where N is the length of the feature vector (number of frames). All speech and non-speech modal intervals, (o_i, o_j) in the feature space would be the pairs of values from \mathbf{o} . If N is the length of **feats** or equivalently \mathbf{o} , total number of possible modal intervals would be $\binom{N}{2} = \frac{N(N-1)}{2}$ that is combinations obtained by choosing two values out of \mathbf{o} vector. Now, for each modal interval (o_i, o_j) we compute the greatest convex minorant, G of empirical distribution, F in $(-\infty, o_i)$ and least concave majorant, H of empirical distribution, F in (o_j, ∞) . Let d_{ij} be the maximum distance between F and curves G, H in modal interval (o_i, o_j) . Then, the DIP is given as

$$DIP = \frac{1}{2} \min \{d_{ij}\}, \quad (1)$$

over all modal interval (o_i, o_j) such that the line segment from $[o_i, F(o_i) + \frac{1}{2}d_{ij}]$ to $[o_j, F(o_j) - \frac{1}{2}d_{ij}]$ lies in the set defined

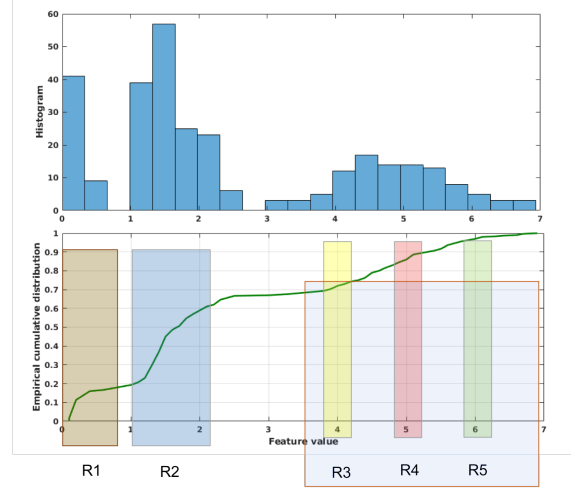


Fig. 1. Illustration of the dip-based clustering technique on synthetic data with five classes, identified with R1 to R5 where three regions R3, R4 and R5 lie close to each other in the feature space.

by

$$\{o, y | o_i \leq o \leq o_j, F(o) - \frac{1}{2}d_{ij} \leq y \leq F(o) + \frac{1}{2}d_{ij}\} \quad (2)$$

The Equation 2 ensures that the greatest convex minorant, modal segment and the least concave majorant together form a unimodal distribution. The Algorithm 1 *computeDip* compute the DIP value, the modal interval and the p-value, p from the significance test.

2.3. Dip-based clustering

We used the dip test recursively to locate the modal intervals that could contain speech or non-speech frames. We explain the proposed clustering approach by looking at Figure 1 and going through the Algorithm 2 *Dip-SAD*. Figure 1 illustrates a simulated scenario showing five categories in the feature space. The top sub-figure shows the histogram of features, while the bottom one shows empirical cumulative distribution. Clearly, the region R3, R4 and R5 lie close to each other in the feature space. On applying the clustering approach described in Algorithm 2 *Dip-SAD*, the first modal interval detected consisted of R3, R4 and R5 (Step 3 in *Dip-SAD*). On recursing again in this interval for each o_j such that $o_L \leq o_j \leq o_U$, we get all the three regions R3, R4 and R5 that forms I_{mid} i.e., middle modal intervals (Step 4). Next, we recurse into the right and left side of the primary interval to find if other segments were present (Step 5). While recursing to the left and right, we included the nearest detected modes from respective left or right region, i.e., for left recursion, region R3 is included in the search region while for right recursion, region R5 is included in the search region (Step 6). Thus, upper limit (u) for left search is minimum among all detected upper limits, i.e., upper limit of region R3. On the

Algorithm 2 *Dip-SAD*

Input: frame-level speech features from an utterance**Output:** speech non-speech labels for each frame

Step 1: Sort the features in ascending order and let $\mathbf{o}=[o_1, o_2, \dots, o_N]$ be the ordered vector, where $o_1 \leq o_2 \leq \dots \leq o_N$. The significance level, α is set to 0.05 for all experiments reported in this paper.

Step 2: $\{o_L, o_U, p\} \leftarrow \text{computeDip}(\mathbf{o})$

Step 3: If $p > \alpha$, then the detected primary modal interval is $[o_L, o_U]$. Else, $[o_1, o_N]$ is primary modal interval.

Step 4: Recurse into the modal interval to find the list I_{mid} of the modal intervals within detected primary mode.

Step 5: Now, we check to the right and left of the primary modal interval recursively and extract additional modes if found.

Step 6: $\{u\} \leftarrow \min_{o_U \in I_{mid}} (o_U), \{l\} \leftarrow \max_{o_L \in I_{mid}} (o_L)$.

Step 7: $p_l \leftarrow \text{computeDip}(\forall o_j : o_j \leq u)$, $p_u \leftarrow \text{computeDip}(\forall o_j : o_j \geq l)$.

Step 8: $I_l \leftarrow$ If $p_l \leq \alpha$, then $\forall o_j : o_j < o_l$ forms a multi-mode segment. We recurse into this interval and return all found modal intervals. Else return ϕ i.e., an empty set.

Step 9: $I_r \leftarrow$ If $p_u \leq \alpha$, then $\forall o_j : o_j > o_u$ forms a multi-mode segment. We recurse into this interval and return all found modal intervals. Else return ϕ i.e., an empty set.

Step 10: The final set of all modal interval is $I_l \cup I_{mid} \cup I_r$.

Step 11: As we knew that combo-SAD features have high positive value for speech and low value for different noises, the cluster with highest average feature value is taken as speech and rest clusters as non-speech. In some instances, where two prominent noise sources were present such as non-stationary background noise and occasional tonal impulsive noise, this approach led to three or more clusters.

other hand, lower limit (l) for right search is chosen as maximum among all lower limits in detected regions, i.e., lower limit of R5 (Step 6). This strategy ensures that the left and right searches will either have unimodal (means same region extended in that direction such as R5 here extends till the end of the right region) or have multi-modalities (means different modes in that direction such as R1 and R2 in left). This is done in Step 6 of the Algorithm 2 *Dip-SAD*. After we have upper limit, u and lower limit, l for left and right searches respectively, we iterate using Algorithm 1 *computeDip* on both regions to get the corresponding p-values, p_l and p_u (Step 7). From the corresponding p-values of such recursions, we conclude unimodality if $p_l > \alpha$ and return empty set ϕ . If $p_l \leq \alpha$, we find the corresponding modal interval and add it to set I_l that is set of modal intervals for left region (Step 8). Similarly, we do for right search (Step 9) to get set I_r that is

set of modal intervals for the right region. The final set of all modal intervals is the union of middle set I_{mid} , left set I_l and right set I_r . The Figure 1 was for illustration of the dip-based clustering approach. For speech activity detection (SAD), at the end of recursive dip tests on detected primary modal interval, left region and right region, we usually get just two, three or four clusters. We found that when there are more than one type of noise in an utterance such as non-stationary background noise, occasional impulsive noise *etc* then each non-speech region with a specific noise-type got clustered separately. From [3], we know that the *Combo* features are relatively large positive values for speech and very small positive or negative values for noise. We leverage this fact in assigning clusters to speech or non-speech class. The cluster with highest average sample value was assigned to speech and rest clusters corresponded to non-speech. This assignment was done automatically on the basis of average feature value for each detected cluster. Authors in [17] also noticed that the *Combo* features for OpenSAD data were significantly tri-modal on some channels and tri-modal GMM helped in gaining improvements in DCF (Section 6.4.1) [17].

3. RESULTS & DISCUSSIONS

We used 40ms windows with a 10ms skip-rate for extracting the *Combo* features from each utterance. The sampling rate for processing the speech data was kept at 8kHz. The *NIST OpenSAD 2015* program was organized to advance the state-of-the-art SAD over extremely degraded communication channels [21]. Six channels namely B, D, E, F, G and H from the DARPA RATS were included in the training set along with the source channel. This data consisted of re-transmitted telephone conversations captured through different communication channels. This data was provided at 16 kHz sampling rate with 16 bit resolution. We downsample the OpenSAD data to 8 kHz for feature extraction and further processing. In this study, we evaluate all channels of the training set as techniques being evaluated are fully unsupervised and parameter-free.

Recently, NIST organized speech analytic technologies evaluation *NIST OpenSAT 2017* [22]. It had three tasks: SAD, key word search, and automatic speech recognition. We evaluated the proposed SAD approach for OpenSAT public safety communications (PSC) data. It contained audio recordings from sofa super store fire (SSSF) dispatcher that occurred on June 18, 2007 in Charleston, South Carolina. The data constitute real fire-response operational data that can not be duplicated through controlled scientific collection [22]. Thus, the data is rich in naturalistic distortions such as (i) land mobile radio transmission effects; (ii) speech under cognitive and physical stress; (iii) varying background noise types and levels *etc* [22]. The data consisted of six audio recordings, each of approximately five minute duration, thus making up a total of 30 minutes of *dev* data. The data were provided as 16-bit signed integer PCM at 8 kHz sampling rate. The *dev* set

Table 1. DCF (%) with two-second collar on all channels of Levantine Arabic (alv) in training set of NIST OpenSAD-2015 data. The scores were averaged over all audio files.

System	alv – B	alv – D	alv – E	alv – F	alv – G	alv – H	alv – src
Combo-SAD	8.54	7.21	6.09	5.60	1.51	6.07	3.02
Proposed	13.21	6.40	5.83	4.19	1.34	3.63	2.68
Relative Improvement (%)	-54.68	11.23	4.27	25.18	11.26	40.20	11.26

Table 2. DCF (%) with two-second collar on all channels of American English (eng) in training set of NIST OpenSAD-2015 data.

System	eng – B	eng – D	eng – E	eng – F	eng – G	eng – H	eng – src
Combo-SAD	9.65	10.32	6.44	5.83	8.18	5.66	4.18
Proposed	10.68	8.18	5.12	2.96	9.30	4.11	6.87
Relative Improvement (%)	-10.67	20.74	20.50	49.23	-13.69	27.38	-64.35

Table 3. DCF (%) with two-second collar on all channels of Urdu (urd) in training set of NIST OpenSAD 2015 data.

System	urd – B	urd – D	urd – E	urd – F	urd – G	urd – H	urd – src
Combo-SAD	7.63	6.98	5.69	5.76	3.73	5.62	3.48
Proposed	5.85	5.51	5.30	5.26	3.67	4.78	4.22
Relative Improvement (%)	23.33	21.06	6.85	8.68	1.61	14.95	-21.26

was shipped with the ground-truth SAD reference labels for evaluation.

The evaluation metric used in NIST OpenSAD-2015 and NIST OpenSAT-2017 was the detection cost function (DCF) given by:

$$DCF = 0.25 * P_{fa} + 0.75 * P_{miss} \quad (3)$$

where P_{fa} is the false alarm rate (non-speech frames detected as speech) and P_{miss} is the miss rate (speech frames detected as non-speech). The DCF values were computed for each audio file and averaged to get the DCF for each channel over three languages in NIST OpenSAD. We incorporated the two-second collar around each speech region in accordance with the NIST OpenSAD 2015 protocol. Table 1, Table 2 and Table 3 shows the comparison of results obtained with the proposed technique using a significance level, $\alpha = 0.05$ and Combo-SAD baseline. The baseline Combo-SAD approach had *Combo* features considered for fitting a two-component GMM. We chose 0.5 weights for both speech and non-speech GMM during threshold selection in baseline [3]. Fixing the weights made the approach parameter-free. Clearly, we can see that the proposed approach led to significant relative gains in DCF as compared to the baseline Combo-SAD except for alv-B, eng-B, eng-G, eng-src, urd-src channels. The Combo-SAD baseline is a model-based technique and it performs well when Combo features are bi-modal. On channels where Combo-SAD is better than the proposed Dip-SAD approach, we found that Combo feature were distinctly bi-modal for majority of the utterances. Overall, we found that the Dip-SAD had reasonable DCF gains over Combo-SAD on many channels. The poor performance of Dip-SAD on some channels

Table 4. DCF with no collar on all audio recording in PSC SSSF dev set from NIST OpenSAT 2017.

Audio name	GMM (%)	Proposed (%)	Relative Improvement(%)
sssf_dev_001	10.04	8.76	12.75
sssf_dev_002	9.25	11.03	-19.24
sssf_dev_003	6.20	5.67	8.55
sssf_dev_004	4.39	4.57	-4.10
sssf_dev_005	6.58	5.13	22.04
sssf_dev_006	8.29	7.88	4.95
Overall	7.46	7.17	3.89

is possibly due to over-clustering of speech into two clusters. In future, we would consider semi-supervised cluster assignments for such cases. Table 4 shows the DCF with *no collar* for all audio recordings in NIST OpenSAT PSC SSSF dev set. We can see that the proposed Dip-SAD approach has overall 3.89% relative improvement in DCF as compared to GMM baseline with same features.

4. CONCLUSIONS

This study leverages Hartigan dip test for unsupervised speech activity detection for scenarios that lack annotations. We used Combo features in proposed clustering approach as these were found to perform well on extremely noisy DARPA RATS data. The proposed approach is deterministic and parameter-free. Results on NIST OpenSAD-2015 data shows proposed approach to be significantly better than the baseline on many channels from three languages. The overall relative improvement in DCF was 3.89% for NIST OpenSAT.

5. REFERENCES

- [1] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Computer Speech & Language*, vol. 47, pp. 132–156, 2018.
- [2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.
- [3] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.
- [4] H. Dubey, A. Sangwan, and J. H. L. Hansen, "A robust diarization system for measuring dominance in peer-led team learning groups," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 319–323.
- [5] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.
- [6] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.
- [7] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.
- [8] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.
- [9] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.
- [10] J. M. Górriz, J. Ramírez, E. W. Lang, and C. G. Puntonet, "Hard C-means clustering for voice activity detection," *Speech Communication*, vol. 48, no. 12, pp. 1638–1649, 2006.
- [11] H. Dubey, A. Sangwan, and J. H. L. Hansen, "Using speech technology for quantifying behavioral characteristics in peer-led team learning sessions," *Computer Speech & Language*, vol. 46, pp. 343–366, 2017.
- [12] 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 *ISCA INTERSPEECH*, 2012, pp. 1969–1972.
- [13] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," in *ISCA INTERSPEECH*, 2013, pp. 3497–3501.
- [14] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *IEEE ICASSP*, 2015, pp. 4500–4504.
- [15] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. L. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra *et al.*, "All for one: feature combination for highly channel-degraded speech activity detection," in *ISCA INTERSPEECH*, 2013, pp. 709–713.
- [16] S. Novotney, D. Karakos, J. Silovsky, and R. Schwartz, "BBN technologies' OpenSAD system," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 8–12.
- [17] M. Graciarena, L. Ferrer, and V. Mitra, "The SRI system for the NIST OpenSAD 2015 speech activity detection evaluation," in *ISCA INTERSPEECH*, 2016, pp. 3673–3677.
- [18] P. M. Hartigan, "Computation of the dip statistic to test for unimodality," *Applied Statistics*, vol. 34, pp. 320–325, 1985.
- [19] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *The Annals of Statistics*, pp. 70–84, 1985.
- [20] S. Maurus and C. Plant, "Skinny-dip: clustering in a sea of noise," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1055–1064.
- [21] "NIST OpenSAD challenge 2015," (Date last accessed 5-July-2016). [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/Open_SAD_Eval_Plan_v10.pdf
- [22] "NIST 2017 pilot speech analytic technologies evaluation, OpenSAT 2017," (Date last accessed 25-Oct-2017). [Online]. Available: https://www.nist.gov/sites/default/files/documents/2017/05/01/nist_2017_pilot_opensat_eval_plan_v2.1_05-01-17.pdf