ROBUST FEATURE CLUSTERING FOR UNSUPERVISED SPEECH ACTIVITY DETECTION

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

{harishchandra.dubey, abhijeet.sangwan, john.hansen}@utdallas.edu

Robust Speech Technologies Lab (RSTL)

Center for Robust Speech Systems (CRSS)

Erik Jonsson School of Engineering & Computer Science

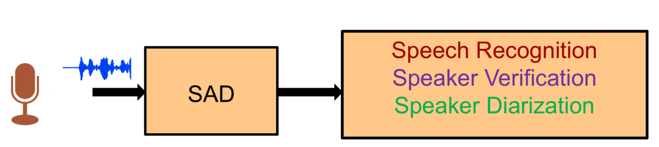
The University of Texas at Dallas

Richardson, Texas 75080-3021, U.S.A.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 15–20 April 2018, Calgary, Alberta, Canada

1 Introduction

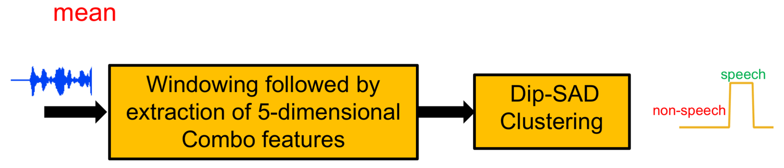
- **Background**
  - Speech activity detection (SAD) is front-end for speech systems
  - Target applications: zero-resource speech processing, low-resource speech systems
- **Proposition**
  - Unsupervised SAD using Hartigan dip test in a recursive-strategy



1. Performance of SAD directly impacts follow-on speech tasks
2. Unsupervised SAD offers flexibility for diverse changing environments

2 Proposed SAD

- **Baseline**
  - Combo features modelled with two-component GMM
  - Component with higher mean correspond to Speech
  - Component with lower mean correspond to Non-speech
  - SAD threshold is convex combination of higher mean and lower mean

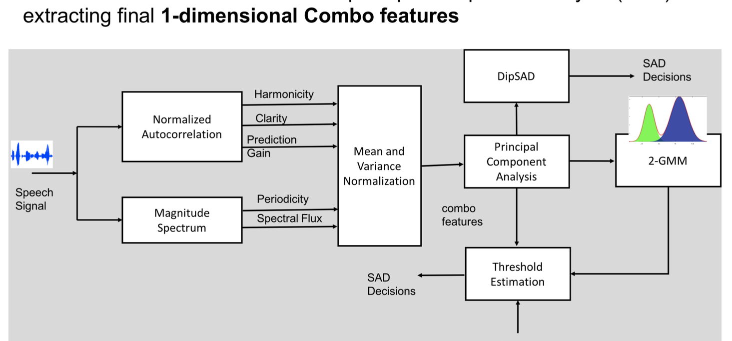


- **Dip-SAD**
  - Combo features considered for recursive Dip-SAD scheme based on Hartigan dip test
  - Parameter-free and deterministic approach

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.

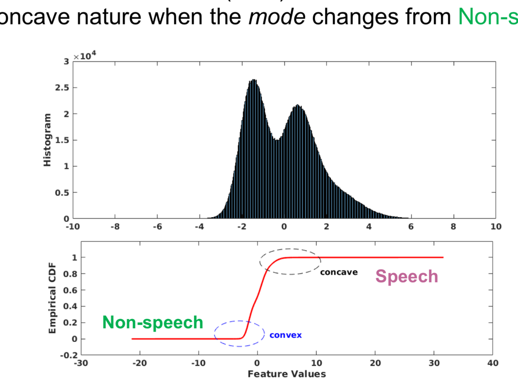
3 Combo Feature

- **Three Steps:**
  - Handcrafted five-dimensional feature-set extracted from time-domain and frequency-domain information
  - Mean and variance normalization performed on each feature dimension
  - Normalized features combined with principal component analysis (PCA) for extracting final **1-dimensional Combo features**



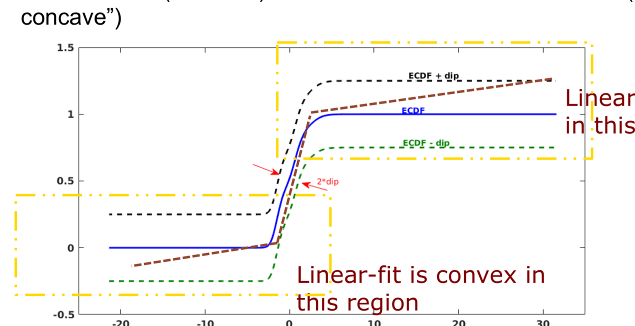
4 Cumulative Distribution

- **Top Figure:**
  - Histogram of Combo features from an utterance (audio recording)
- **Bottom Figure:**
  - Cumulative distribution (CDF) of Combo features follow the convex-then-concave nature when the *mode* changes from Non-speech to Speech



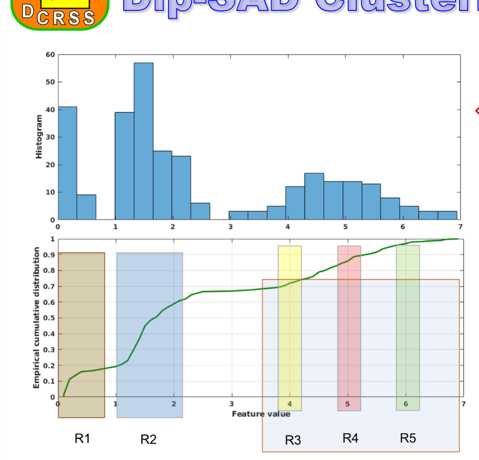
5 Dip in Cumulative Distribution

- Distribution is unimodal if its cumulative distribution takes a convex form up to its mode/modal interval and a concave form after it.
- Dip statistic, dip ∈ (0, 1/4], is defined as the minimum achievable vertical offset for two copies of ECDF (**one above**, **one below**, shown in dashed lines) such that **linear-fit** (of ECDF) does not violate its unimodal rules (i.e., "convex then concave")



Farther distribution move away from unimodality, larger the corresponding dip

6 Dip-SAD Clustering: illustration



- **Clustering examples with 5 acoustic classes:**
  - First iteration gives: [R3,R5] and [R1,R2]
  - Second iteration within [R3,R5] gives R3,R4 and R5
  - Iterating in [R1,R2] gives R1 and R2

1. We included nearest region in left-search and right-search
2. Dip-SAD on a sorted feature-vector requires O(N) operations in the worst case; where N is number of frames

7 Computing Dip

Algorithm 1 *computeDip*

Input: speech features were sorted in ascending order i.e., $\mathbf{o}$=[o_1, o_2, ...,o_N] where $o_1 \leq o_2 \leq ... \leq o_N$.
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] [2]. Let the points of contact with F are respectively, $g_1, g_2, .., g_k$ (for G) and $h_1, h_2, .., h_m$ (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 {

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

}, 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.

8 Proposed Dip-SAD

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, ...,o_N] be the ordered vector, where $o_1 \leq o_2 \leq ... \leq o_N$. The significance level, α is set to 0.05 for all experiments reported in this paper.
Step 2: $\{o_L, o_U, p\} \leftarrow computeDip(\mathbf{o})$
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: $p_l \leftarrow computeDip(\ \forall o_j : o_j \leq u)$, $p_u \leftarrow computeDip(\ \forall o_j : o_j \geq l)$.
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 $I_l \bigcup I_{mid} \bigcup I_r$.
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.

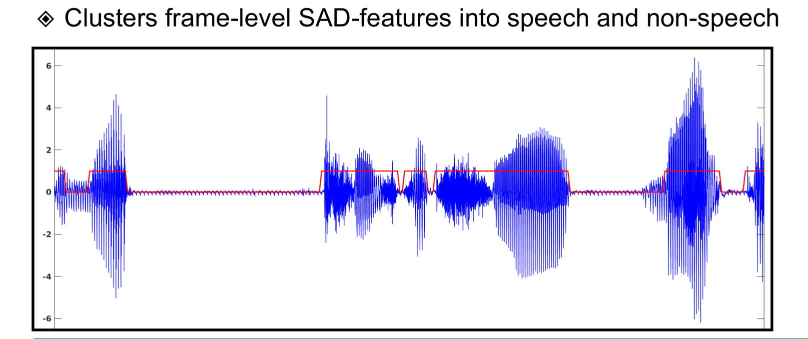
9 NIST OpenSAD-2015

- **Data Stats:**
  - Re-transmitted telephone conversations through six channels B, D, E, F, G and H
  - Subset of DARPA RATS: extremely degraded audio
  - Only OpenSAD training data used in our studies

Channel	Frequency Band	Modulation Type
B	UHF	Narrow-band FM
D	HF	Single side-band AM
E	VHF	Narrow-band FM
F	UHF	Frequency-hopping spread-spectrum
G	UHF	Wide-band FM
H	HF	AM

10 Dip-SAD illustration

- **Specification:**
  - Dip-SAD process all features from a single utterance at a time
  - Clusters frame-level SAD-features into speech and non-speech



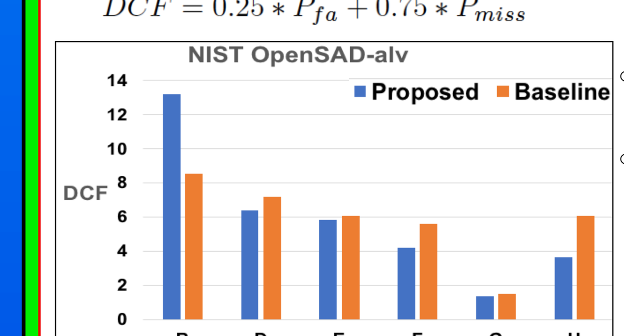
1. Dip-SAD effective for both short and long bursts of speech vocalizations
2. Cluster with highest average feature-value is considered speech and rest non-speech

11 NIST OpenSAD Results-I

- **Levantine Arabic (alv):** DCF with two-second collar

$$DCF = 0.25 * P_{fa} + 0.75 * P_{miss}$$

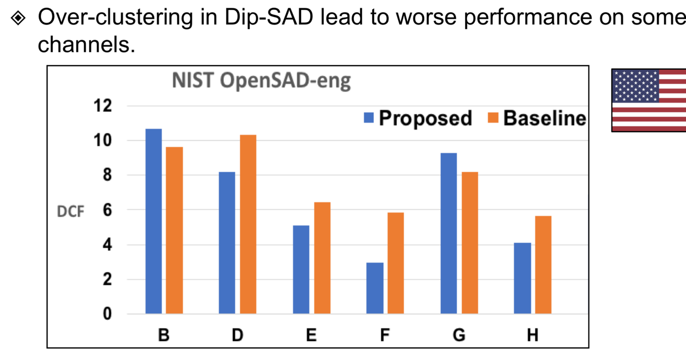
- P_{fa} is the false alarm rate (non-speech frames detected as speech)
- P_{miss} is the miss rate (speech frames detected as non-speech)
- DCF is detection cost function



1. Dip-SAD effective on extremely noisy channels
2. Relative improvements: B (-54.68%); D (+11.23%), E (+4.27%), F (+25.18%), G (+11.26%), H (+40.20%)

12 NIST OpenSAD Results-II

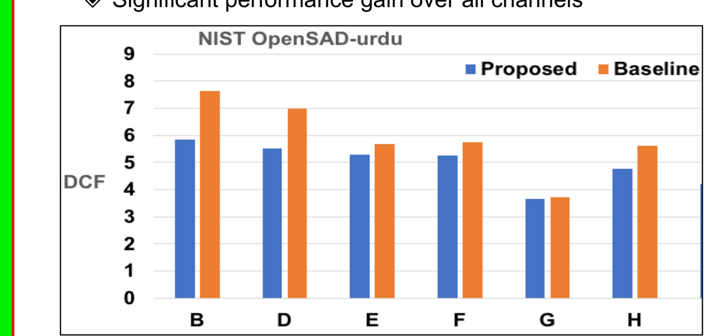
- **American English (eng):**
  - Significant improvements over channels D, E, F and H.
  - Over-clustering in Dip-SAD lead to worse performance on some channels.



1. Dip-SAD effective on extremely noisy channels
2. Relative improvements: B (-10.67%) ; D (+20.74%), E (+20.50%), F (+49.23%), G (-13.69%), H (+27.38%)

13 NIST OpenSAD Results-III

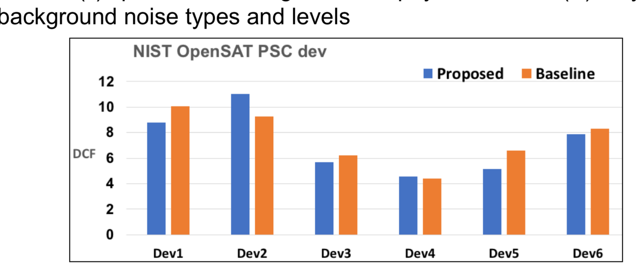
- **Urdu:**
  - Significant performance gain over all channels



1. Dip-SAD effective on extremely noisy channels
2. Relative improvements: B (+23.33%) ; D (+21.06%), E (+6.85%), F (+8.68%), G (+1.61%), H (+14.95%)

14 NIST OpenSAT-2017

- **Public safety communications (PSC): dev data with 30 min. audio**
  - Audio recordings from sofa super store fire dispatcher - Charleston, South Carolina, USA.
  - 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



1. Dip-SAD effective on extremely audio recordings
2. Relative improvements: Dev1 (+12.75%) ; Dev2 (-19.24%), Dev3 (+8.55%), Dev4 (-4.10%), Dev5 (+22.04%), Dev6 (+4.95%)

15 Conclusions & Summary

- **Outcomes:**
  - DipSAD is based on the **geometry of cumulative distribution**
  - Deterministic and parameter-free approach leveraging Hartigan dip test
  - Useful for zero-resource scenarios without SAD transcripts
  - DipSAD significantly better than baseline GMM on NIST OpenSAD-2015
  - Overall relative DCF improvement for NIST OpenSAT-2017: **+3.89%**
  - Over-clustering in DipSAD can lead to poor performance over some channels due to ambiguous assignment of clusters to speech/non-speech
- **Future Work:**
  - Leveraging knowledge of features in accurate assignment of clusters to Non-speech/Speech classes can improve accuracy for non-binary cases (i.e., 3, 4, or more clusters)
  - Incorporating multi-dimensional features in Dip-SAD recursions is likely to lend further robustness and accuracy
  - Strategies for enforcing only binary clustering into Non-speech/speech classes

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.