

Detection of Glottal Closure Instants from Raw Speech using Convolutional Neural Networks



Mohit Goyal*, Varun Srivastava*, Prathosh AP

Indian Institute of Technology, Delhi, India

*Equal Contribution

Introduction

- Glottal Closure Instants (GCIs) - **Temporal locations** of significant excitation to the vocal tract occurring during the production of voiced speech.
- Most of the existing approaches for GCI detection adopt a two-stage approach
 - Transformation** of speech signal into a representative signal where GCIs are localized better,
 - Extraction** of GCIs using the representative signal obtained in first stage.
- The former stage uses signal processing techniques based on the principles of speech production
- The latter stage typically uses dynamic-programming and peak-picking.
- Thus these methods remain task-specific and rely on the methods used for representative signal extraction.
- In this work, GCI detection is cast as a **supervised multi-task learning** problem solved using a deep convolutional neural network jointly optimizing a **classification and regression cost**. The learning capability is demonstrated with several experiments on standard datasets.

Methodology - Data Generation

The GCI detection problem is formulated using a block processing approach shown as follows:

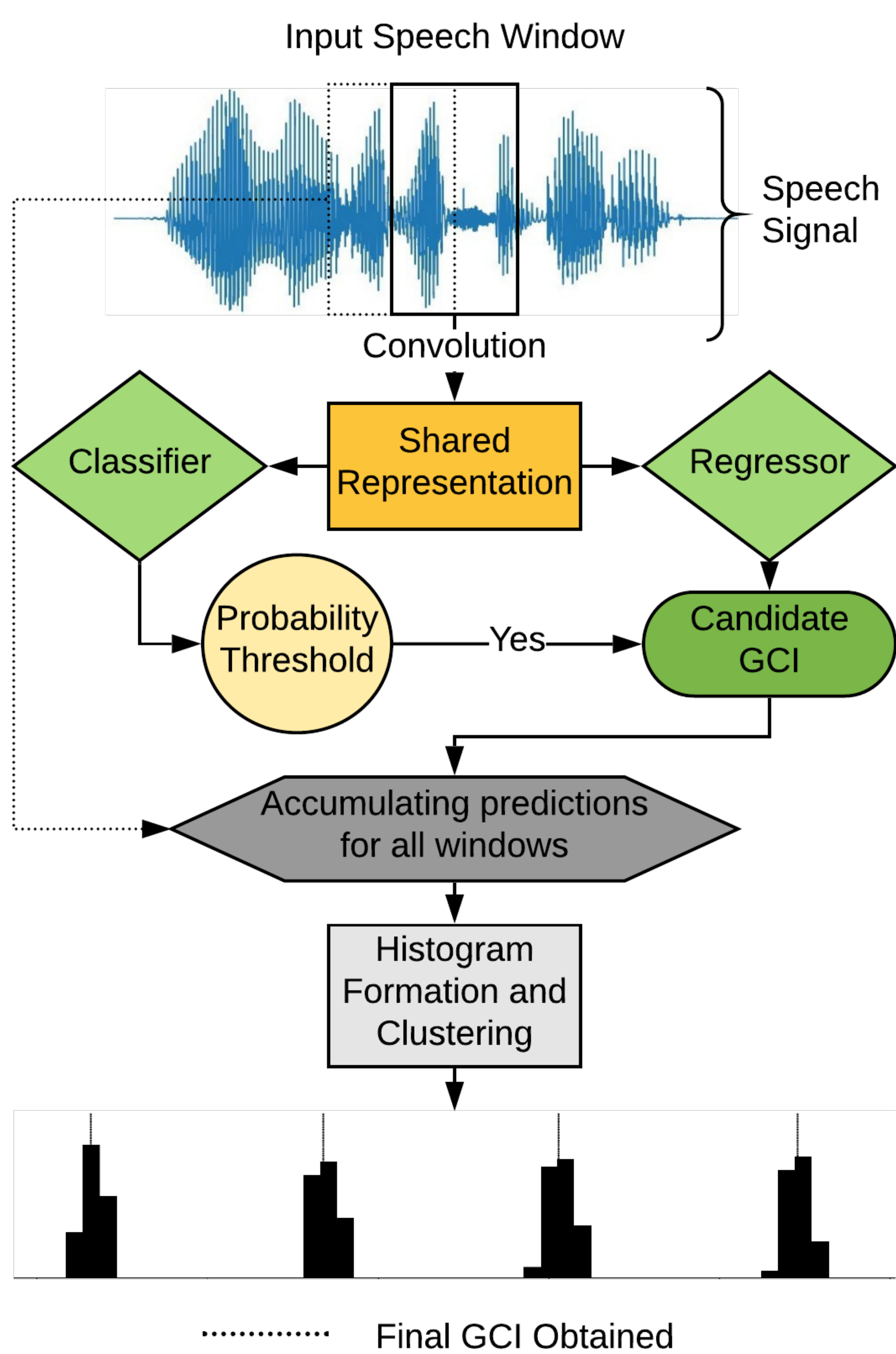
- Divide speech signal into detection windows of length 2 ms (minimum pitch period) - w_d .
 $\Rightarrow$ Each detection window w_d contains *atmost* one GCI.
- Concatenate a context window (1.5 ms) on both sides of w_d generating the input frame - w_i
- Denote presence of GCI in w_d as $t_c = 1$ else 0. If present, denote relative position of GCI within a detection window as t_r where $0 \leq t_r < w_d$.

Proposed Method

In the proposed algorithm, we train a CNN (Convolutional Neural Network) to perform simultaneous regression and classification on each detection window. As a result of jointly learning a shared representation for both tasks, total number of learnable parameters are reduced significantly and an improvement in learning efficiency is observed.

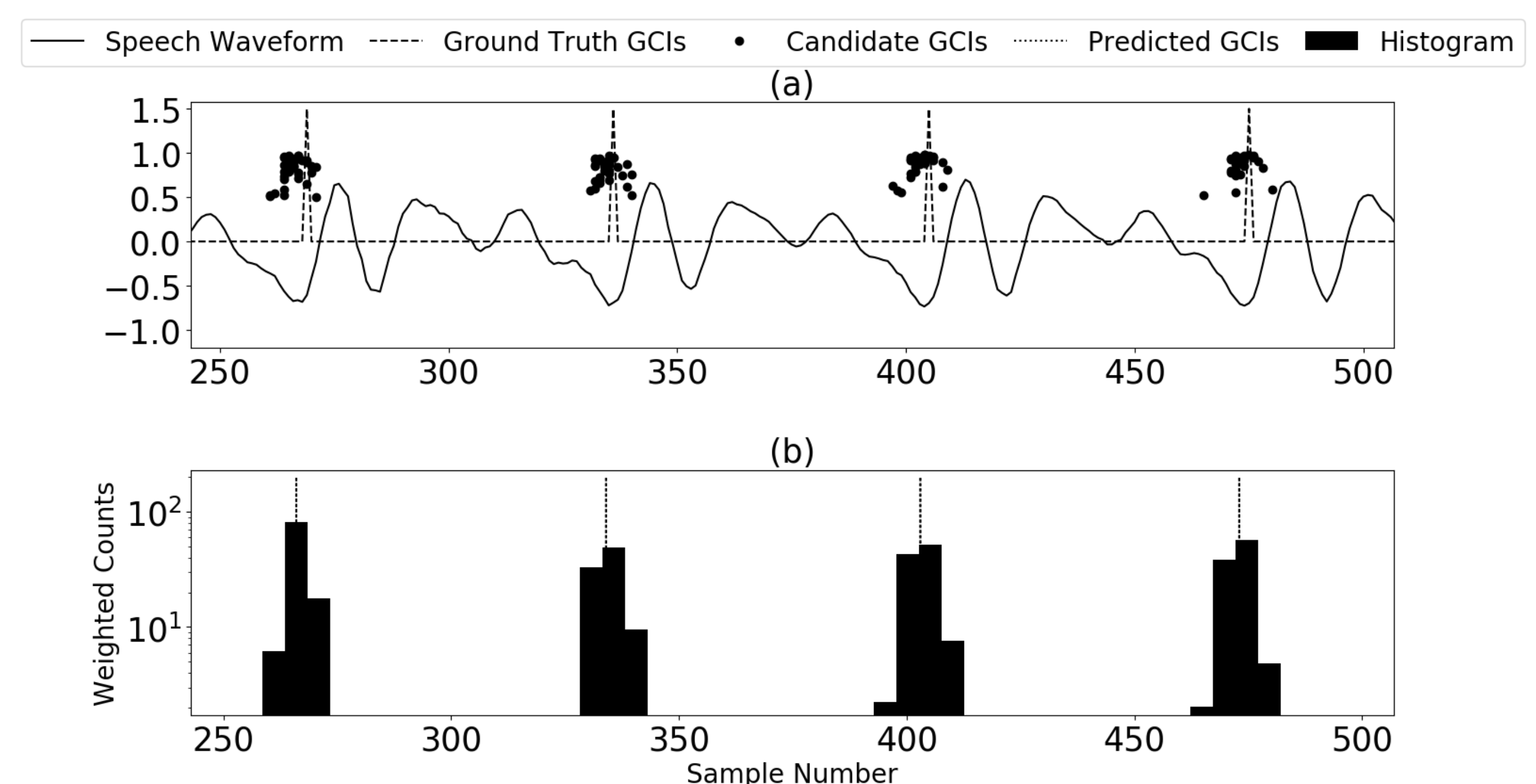
Objective Function: During training, the network predicts the probability (y_c) of observing a GCI and its location (y_r) within the corresponding detection window. Thus, for the j^{th} input frame, we minimize the following multi task loss function.

$$\mathcal{L}^j = t_c^j \log(y_c^j) + (1 - t_c^j) \log(1 - y_c^j) + t_r^j * (t_r^j - y_r^j)^2$$



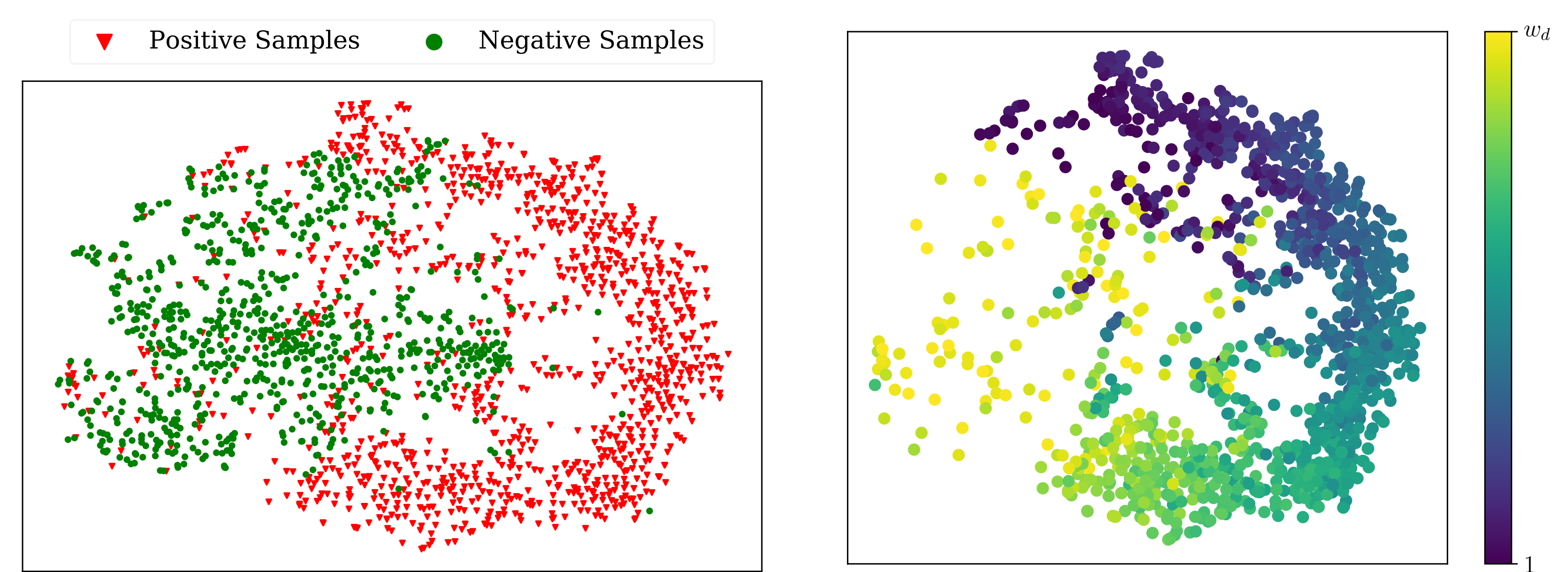
Flowchart depicting the inference procedure given an input speech frame

Illustration for Histogram Clustering



(a) a segment of speech with four ground truth GCIs along with the corresponding candidate GCIs (height represents confidence in detection), (b) weighted histograms shown along with the means of the local-groups (dotted lines) which form the final predictions.

Visualizing Shared Representation



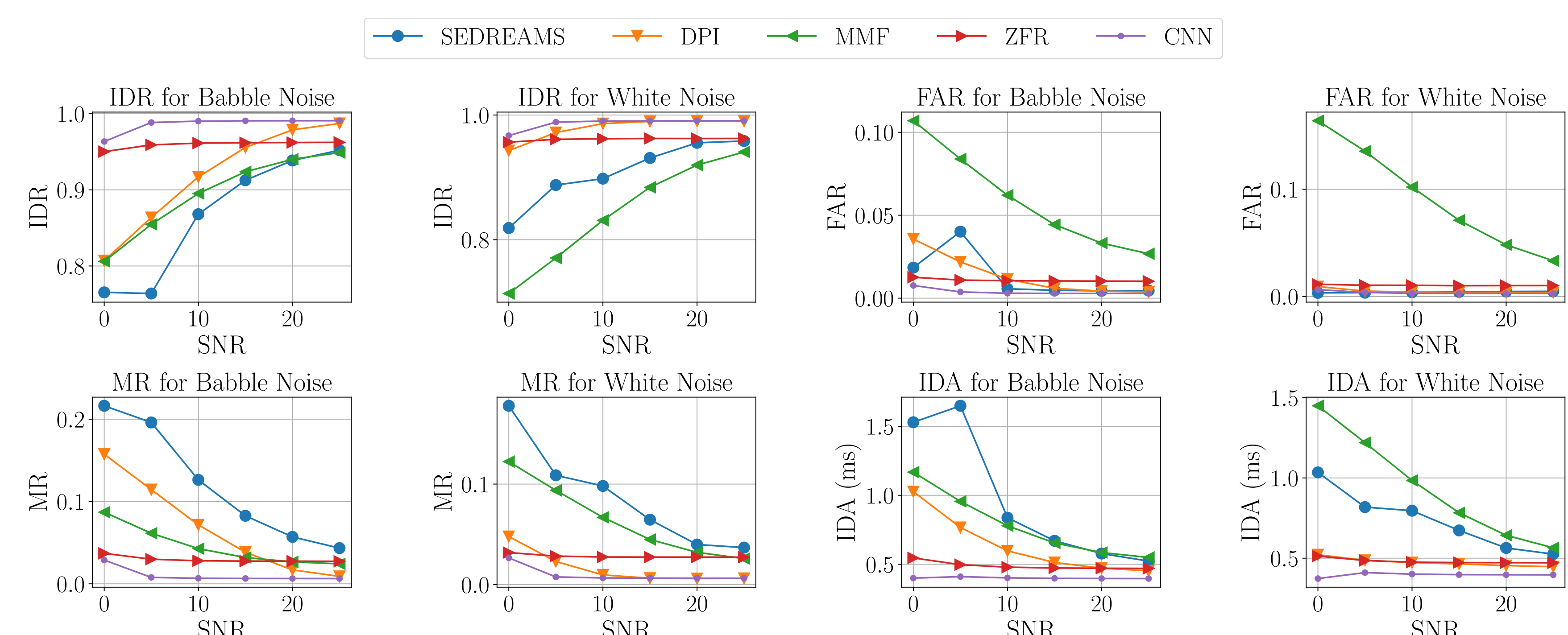
Visualizing shared representation (extracted from CNN) with the help of t-SNE using
(a) classification labels. (b) regression values corresponding to color gradient.

Results

Datasets: CMU dataset—JMK (Canadian Male), BDL (US Male), SLT (US Female)

Metrics: IDR (Identification Rate), MR (Miss Rate), FAR (False Alarm Rate), IDA (Identification Accuracy)

Algorithms chosen for Comparison: Zero-frequency resonator (ZFR)¹, Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS)², Dynamic Plosion Index (DPI)³ and Micro-canonical Multi-scale Formalism (MMF)⁴.



Conclusion

In this project, we propose a novel data driven method for GCI detection from raw speech using multi-task supervised learning. Leveraging on the advantages of multi task learning, we obtained superior performance on noisy speech and performance at par on clean speech compared to state of the art algorithms. While we conducted experiments on speakers of a single dataset, a larger study focusing on multiple datasets would help further establish the accuracy and efficacy of our method. Future work could also explore extending our formulation to detecting temporal events in general time series data, as well as performing real time computation using causal processing of the input speech frame.

References

- [1] K. S. R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, Lang. Process.*, 16, no. 8:1602–1613, Nov. 2008.
- [2] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.
- [3] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan. Epoch extraction based on integrated linear prediction residual using plosion index. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2471–2480, Dec 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2273717.
- [4] Vahid Khanagha, Khalid Daoudi, and Hussein M Yahia. Detection of glottal closure instants based on the microcanonical multiscale formalism. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(12):1941–1950, 2014.