# DeepTalk: Vocal Style Encoding for Speaker Recognition and Speech Synthesis


Anurag Chowdhury[1], Arun Ross[1], Prabu David[2]

1 - Department of Computer Science Engineering, 2 - College of Communication Arts and Sciences

MICHIGAN STATE UNIVERSITY


## Introduction

Automatic speaker recognition algorithms typically characterize speech audio using short-term spectral features that encode the physiological and anatomical aspects of speech production. Such algorithms do not fully capitalize on speaker-dependent characteristics present in behavioral speech features.

In this work, we:

1) Develop a vocal-style encoder called DeepTalk for capturing speaker-dependent behavioral speech characteristics
2) Combine DeepTalk with physiological speech feature-based speaker recognition methods to improve speaker recognition performance in challenging audio conditions
3) Integrate DeepTalk into a Text-To-Speech (TTS) synthesizer to generate synthetic speech audios for evaluating the fidelity of DeepTalk-based vocal style features

## DeepTalk

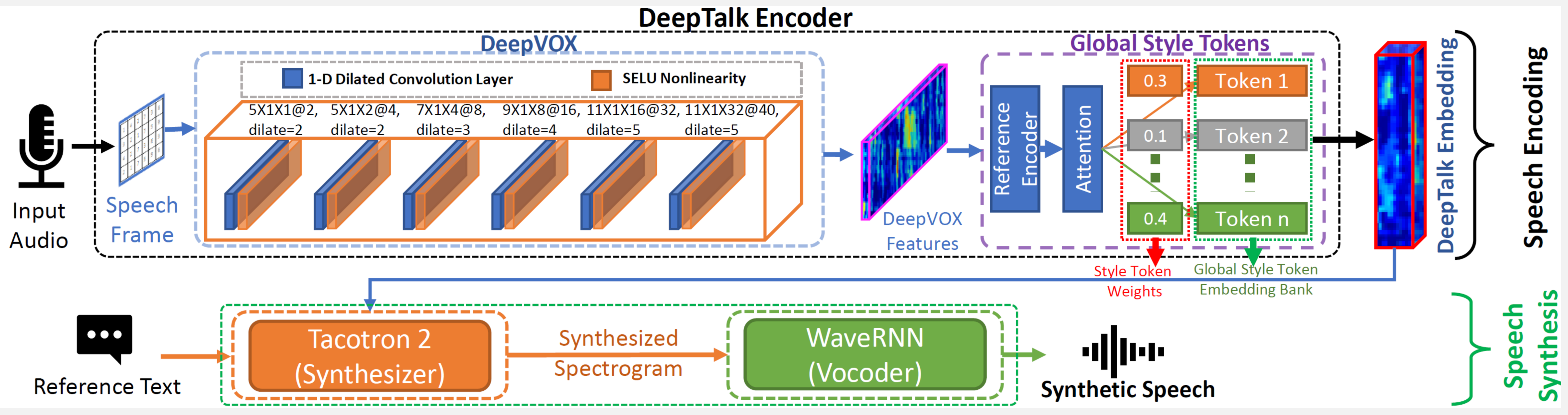



Figure 1: A visual representation of the proposed DeepTalk-based speech encoding and speech synthesis framework

In this work, we develop a speech encoder called DeepTalk, to capture behavioral speech characteristics directly from raw speech audio without any word- or frame-level annotations. The DeepTalk architecture (Fig. 1) consists of separate speech encoding and speech synthesis branches.

- Speech Encoding: The speech encoding branch feeds a raw input audio into a DeepVOX[1] network to extract short-term speech features, called DeepVOX features. DeepVOX is a 1D-CNN based speech filterbank that extracts speaker-dependent speech features directly from raw speech audio. DeepVOX features are then fed to a Global Style Token (GST)-based[2] prosody embedding network to extract the DeepTalk embedding.

- Speech Synthesis: The speech synthesis branch feeds the DeepTalk embedding and a reference text into a Tacotron2-based synthesizer[3] to generate a Mel spectrogram, which is then converted to the synthetic speech waveform using a WaveRNN-based neural vocoder[4]

Experimental results show the efficacy of the DeepTalk embedding for performing both speaker recognition and speech synthesis, as compared to baseline methods.

## Datasets and Experiments

Following experiments are performed in this work:

- **xVector-PLDA**[8], **i-vector-PLDA** [9], and **1D-Triplet-CNN** [10] methods were used to establish **baseline** physiological speaker verification performance.
- **DeepTalk** is used to perform **vocal-style feature-based** speaker verification experiments
- The DeepTalk and baseline methods are **combined** at a weighted score level to evaluate the speaker recognition benefits of combining behavioral and physiological speech features.

**VoxCeleb2 [5]**

**Number of Speakers:**
5,994 in training set
118 in test set

**Type of Speech Data:**
Interview Speech

**NIST SRE 2008 [6]**

**Number of Speakers:**
1336 in training set
200 in test set

**Type of Speech Data:**
Phone call and Interview Speech

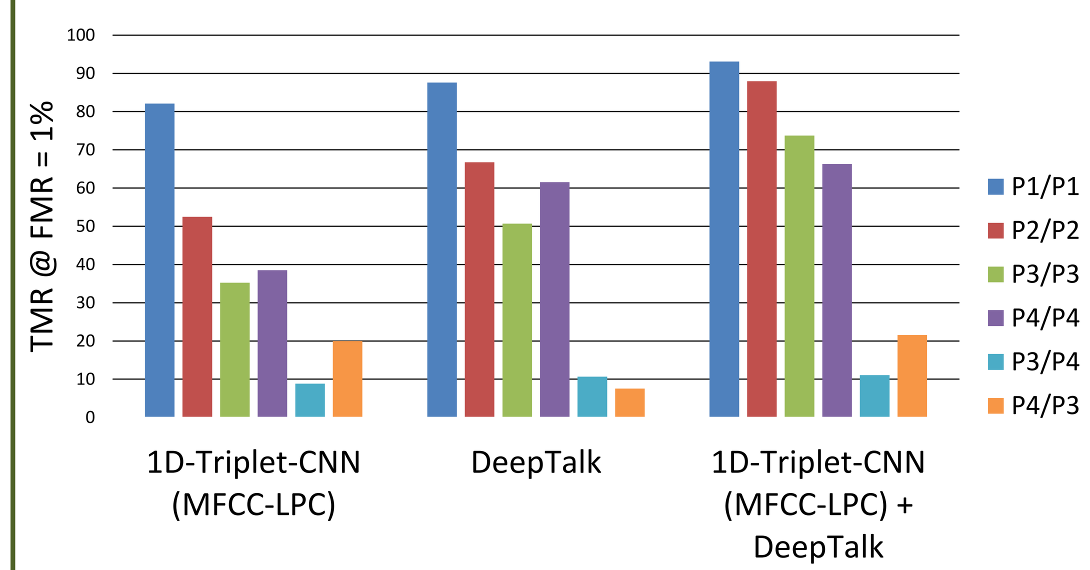
**NOISEX-92 [7]**

**Noise dataset:**
Airplane (F16) Noise
Babble Noise

Figure 2: The above datasets were used for performing the experiments in this work

## Results and Analysis

### Speaker Recognition Results




Train / Test Data:

P1: VoxCeleb2

P2: NIST SRE 2008

P3: NIST SRE 2008 + Babble

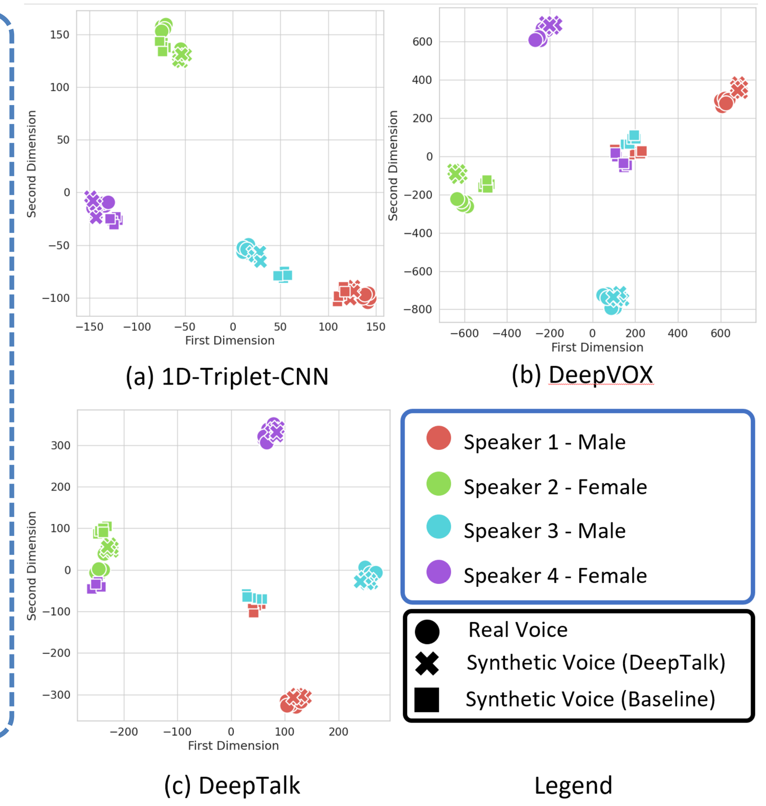
P4: NIST SRE 2008 + F16

Score level fusion of DeepTalk with:

1. 1D-Triplet-CNN(MFCC-LPC) improves TMR@FMR=1% by 19.43%
2. iVector-PLDA improves TMR@FMR=1% by 24.67%
3. xVector-PLDA improves TMR@FMR=1% by 24.24%

### Speech Synthesis Results

- t-SNE plots of the speech embeddings of real and synthetic voice samples of four different speakers, extracted by three different speech encoders.
- DeepTalk's synthetic speech is embedded much closer to the real speech by all the speech encoders, as compared to the baseline synthetic speech.





**Summary**

- Behavioral speech features extracted by DeepTalk method outperform majority of physiological speech feature-based speaker verification methods
- Score-level fusion of DeepTalk with physiological speech feature-based speaker recognition methods further improve the speaker verification performance in majority of the experiments across all the methods
- DeepTalk-synthesized speech is judged near-identical to real speech by SOTA speaker recognition methods, demonstrating DeepTalk's efficacy at vocal style modeling


**Future Work**

- We plan to extend our work towards combining physiological and behavioral speech characteristics at feature-level in a single end-to-end network architecture for further improving the speaker recognition performance.


## Acknowledgement

We thank Susi Elkins of WKAR Public Media for providing us with high-quality audio samples for finetuning the DeepTalk method. Portions of this work were funded by the National Association of Broadcasters.


## References


[1] Chowdhury, Anurag, and Arun Ross. "DeepVOX: Discovering Features from Raw Audio for Speaker Recognition in Degraded Audio Signals." *arXiv preprint arXiv:2008.11668* (2020).
[2] Wang, Yuxuan et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis." In International Conference on Machine Learning, 2018
[3] Shen et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." In IEEE ICASSP, 2018.
[4] Kalchbrenner, et al. "Efficient Neural Audio Synthesis." In *International Conference on Machine Learning*, 2018
[5] Chung, Joon Son et al. "Voxceleb2: Deep speaker recognition." In INTERSPEECH 2018.
[6] "2008 NIST speaker recognition evaluation trainingset part 2 ldc2011s07,"https://catalog.ldc.upenn.edu/LDC2011S05, Accessed: 2018-03-06.
[7] Andrew Varga and Herman JM Steeneken. "Assessmentfor automatic speech recognition: II. NOISEX-92: adatabase and an experiment to study the effect of additive noise on speech recognition systems," Speech communication, 1993.