MICHIGAN STATE UNIVERSITY

DeepTalk: Vocal Style Encoding for Speaker Recognition and Speech Synthesis

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

1 - Department of Computer Science and Engineering

2 - College of Communication Arts and Sciences

Michigan State University

iPRoBe

Voice Biometrics

Voice Biometric

Physical Traits

Short-term spectral features:
Model the Vocal cords

Vocal Source Features:
Model the Lungs, Trachea etc.

Highly discriminative features

Good for speaker recognition

Behavioral Traits

Prosodic features:
Pitch, loudness, duration, timbre

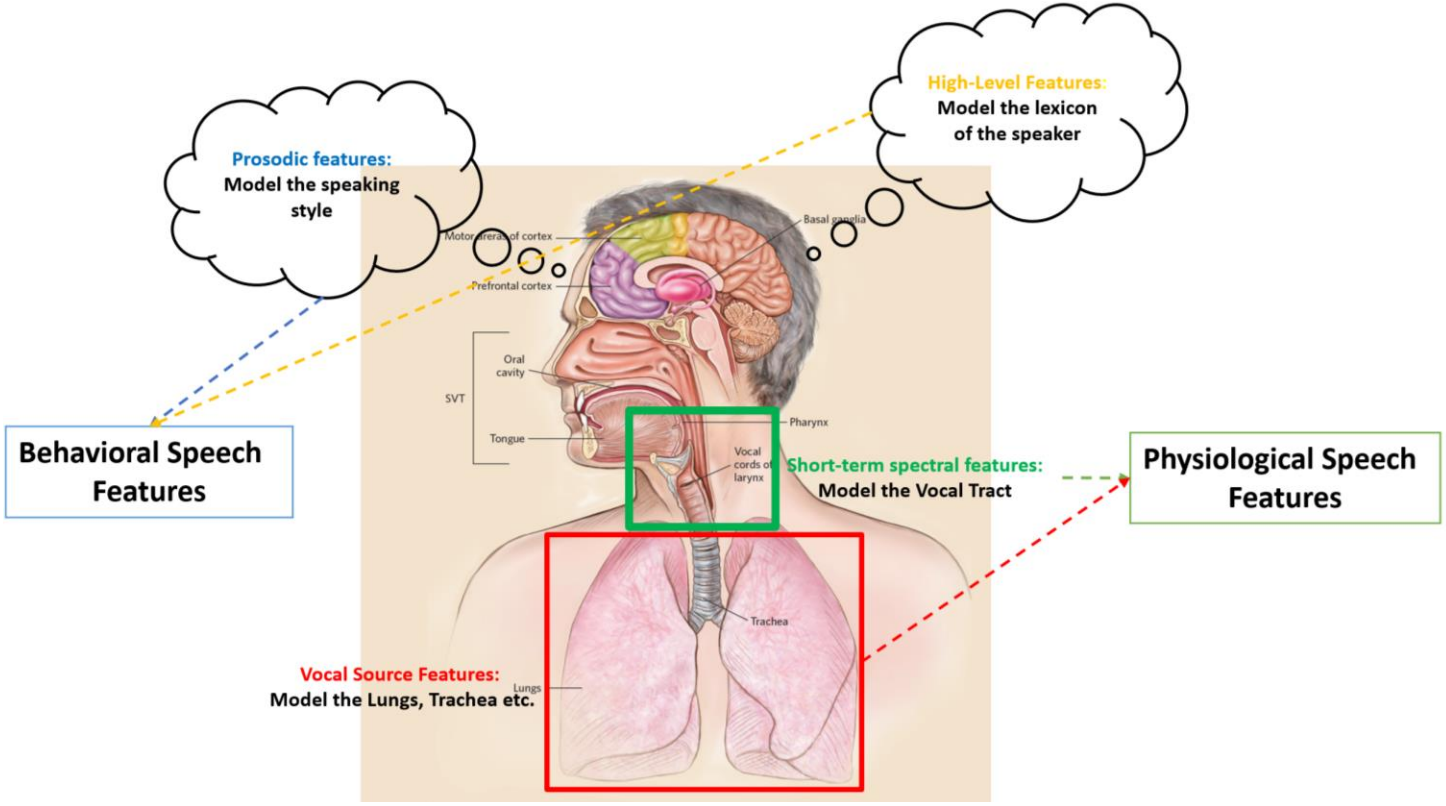
High Level Features:
Model the lexicon of the speaker

Encodes Speaking Style

Important for realistic speech synthesis

Role of Speaking Style in Voice Biometrics

- Majority of speaker recognition methods only use physical traits of human voice
- The volatile nature of speaking style makes it difficult to model
- Speaking style varies with emotional state, language, content and context of speech [1]
- Speaking style contains complementary speaker-dependent characteristics [2]
- Behavioral traits can be combined with physical traits to improve speaker recognition performance [2]



[1] Mary, Leena. "Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition." In *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*, pp. 1-22. Springer, Cham, 2019.
[2] Andre G. Adami, Radu Mihaescu, Douglas A.Reynolds, and John J. Godfrey, “Modeling prosodicdynamics for speaker recognition,” in IEEE ICASSP, 2003.

Contributions of this work

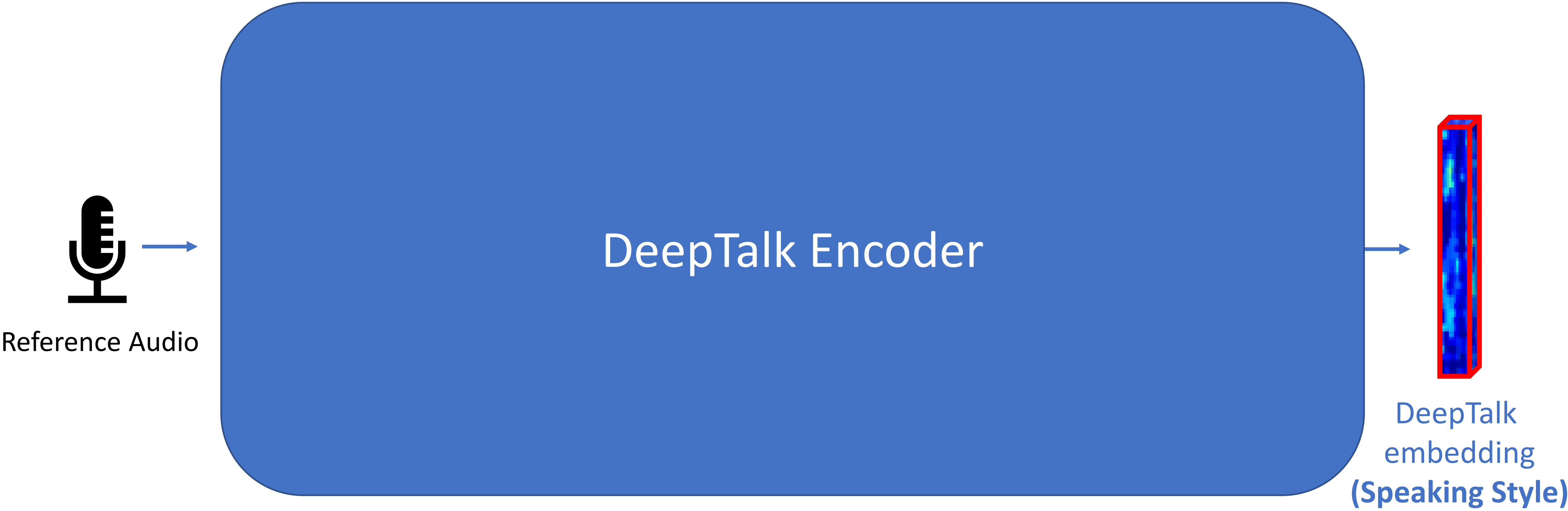
1) Develop a vocal-style encoder called DeepTalk for capturing speaker-dependent behavioral speech characteristics

2) Combine DeepTalk with physiological speech feature-based speaker recognition methods to improve speaker recognition performance in challenging audio conditions

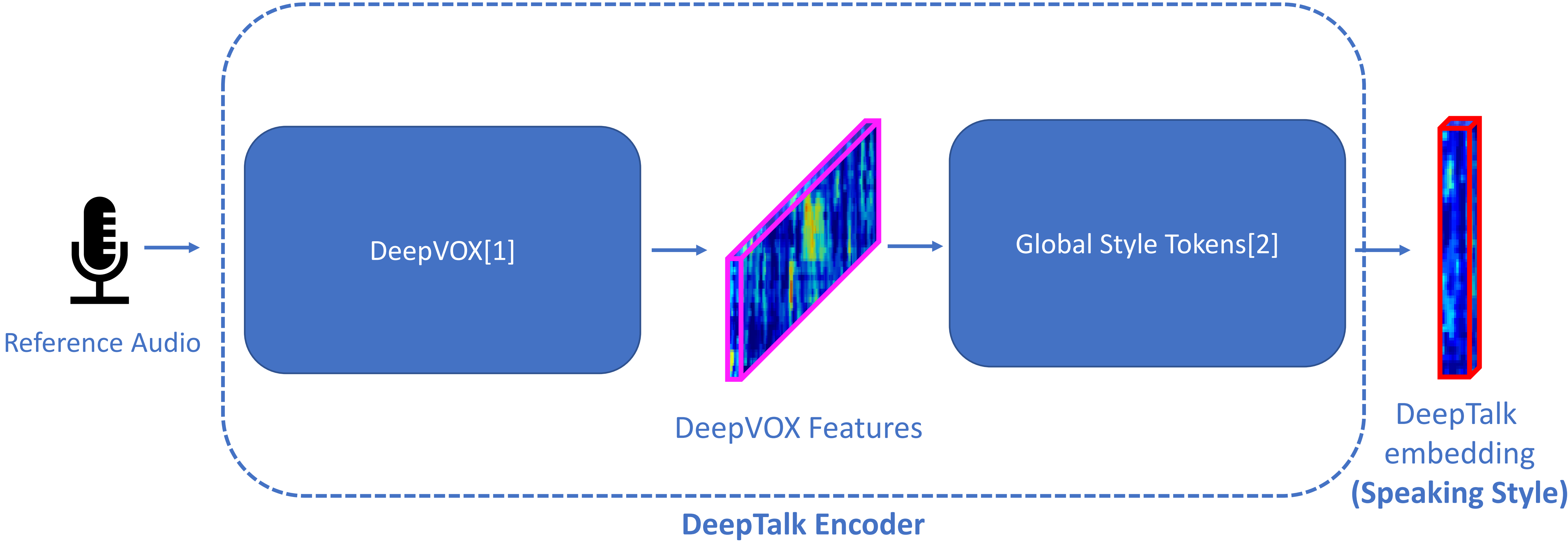
3) Integrate DeepTalk into a Text-To-Speech (TTS) synthesizer to generate synthetic speech audios for evaluating the fidelity of DeepTalk-based vocal style features

DeepTalk:
Vocal Style Encoding for Speaker Recognition

DeepTalk Encoder Design



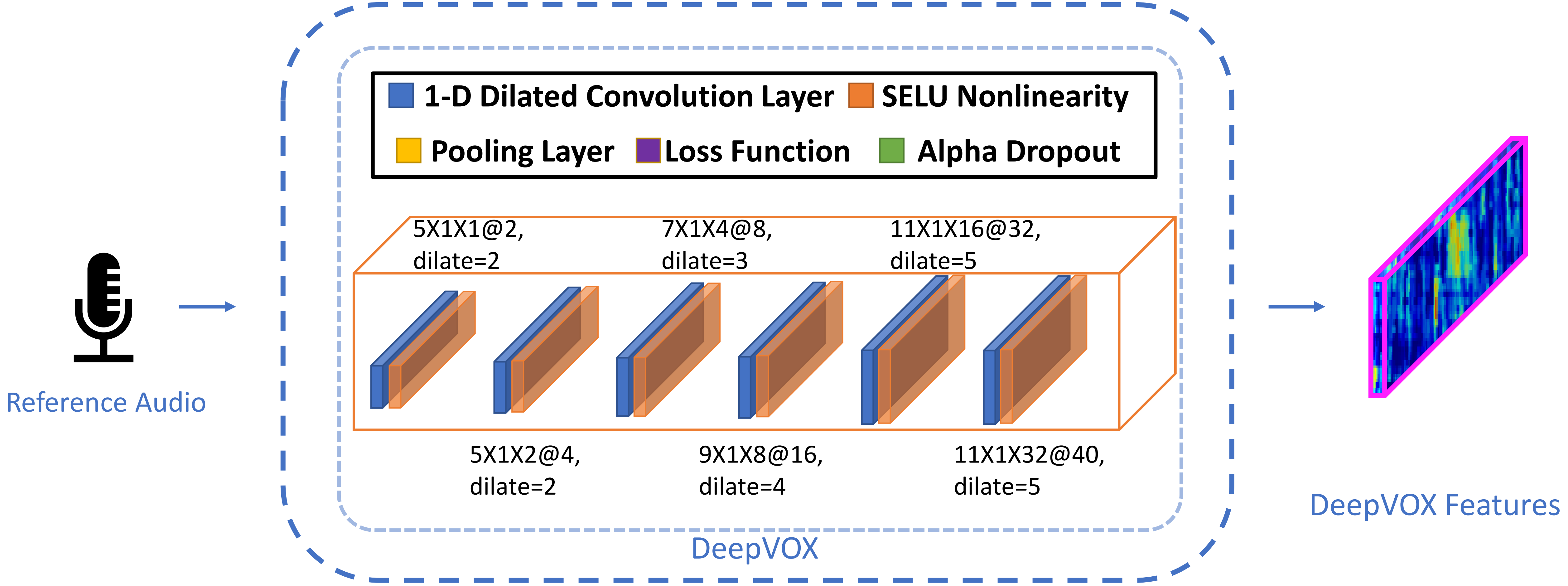
DeepTalk Encoder Design



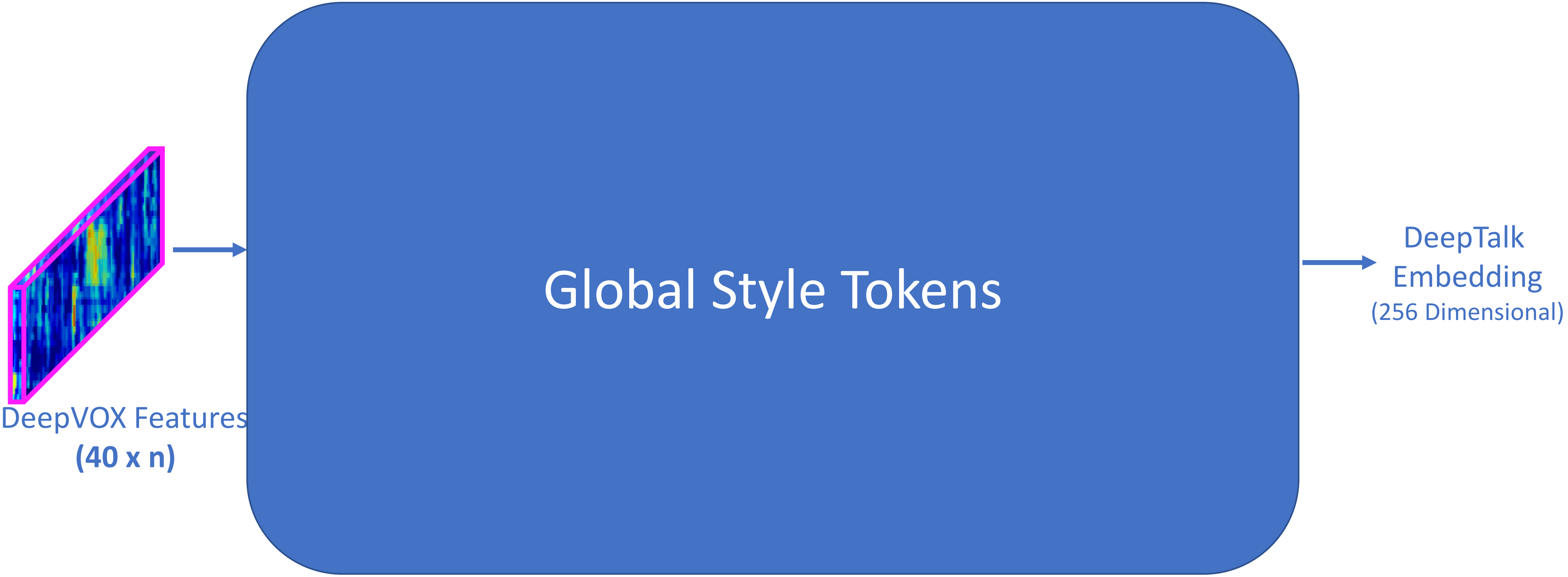
[1] Chowdhury, Anurag, and Arun Ross. "DeepVOX: Discovering Features from Raw Audio for Speaker Recognition in Degraded Audio Signals." *arXiv preprint arXiv:2008.11668* (2020).
[2] Wang, Yuxuan et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis." In *International Conference on Machine Learning*, pp. 5180-5189. 2018.

DeepTalk Encoder Design

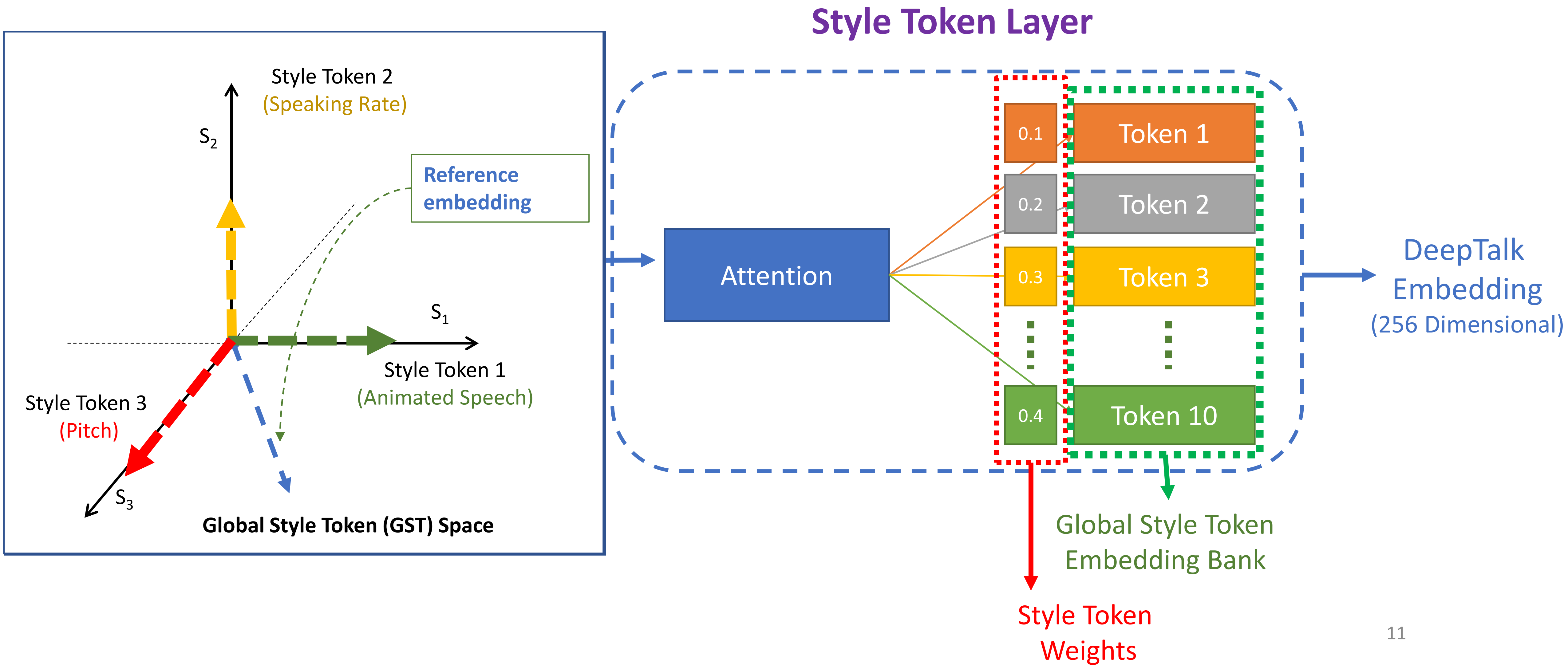
DeepTalk Encoder Design: DeepVOX based speech feature extraction



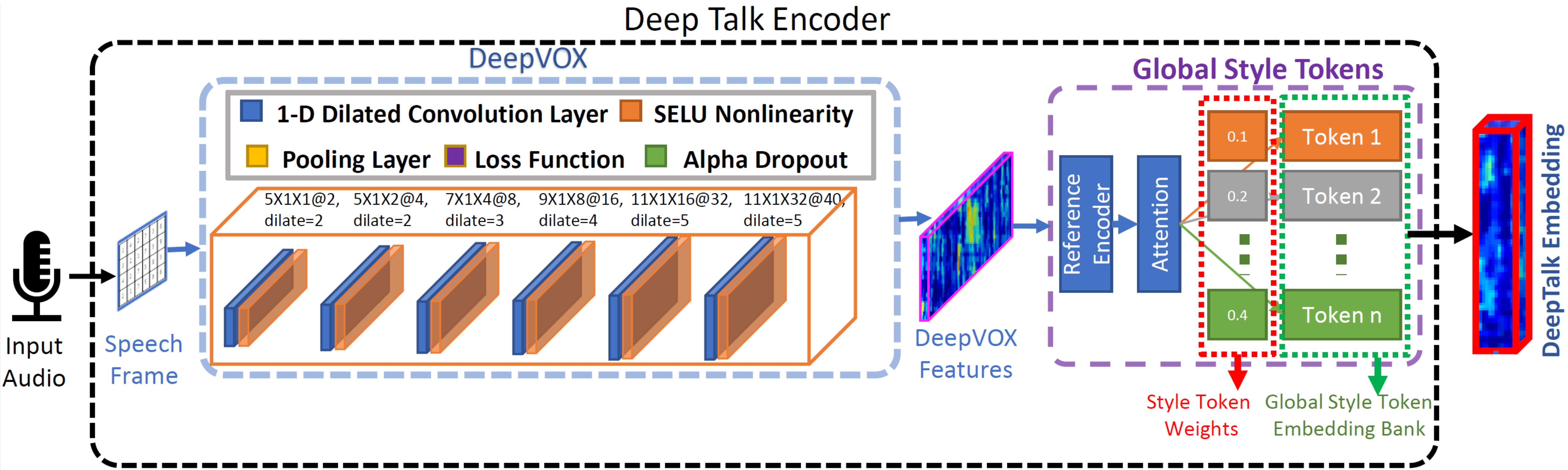
DeepTalk Encoder:
Global Style Token(GST) based prosody embedding



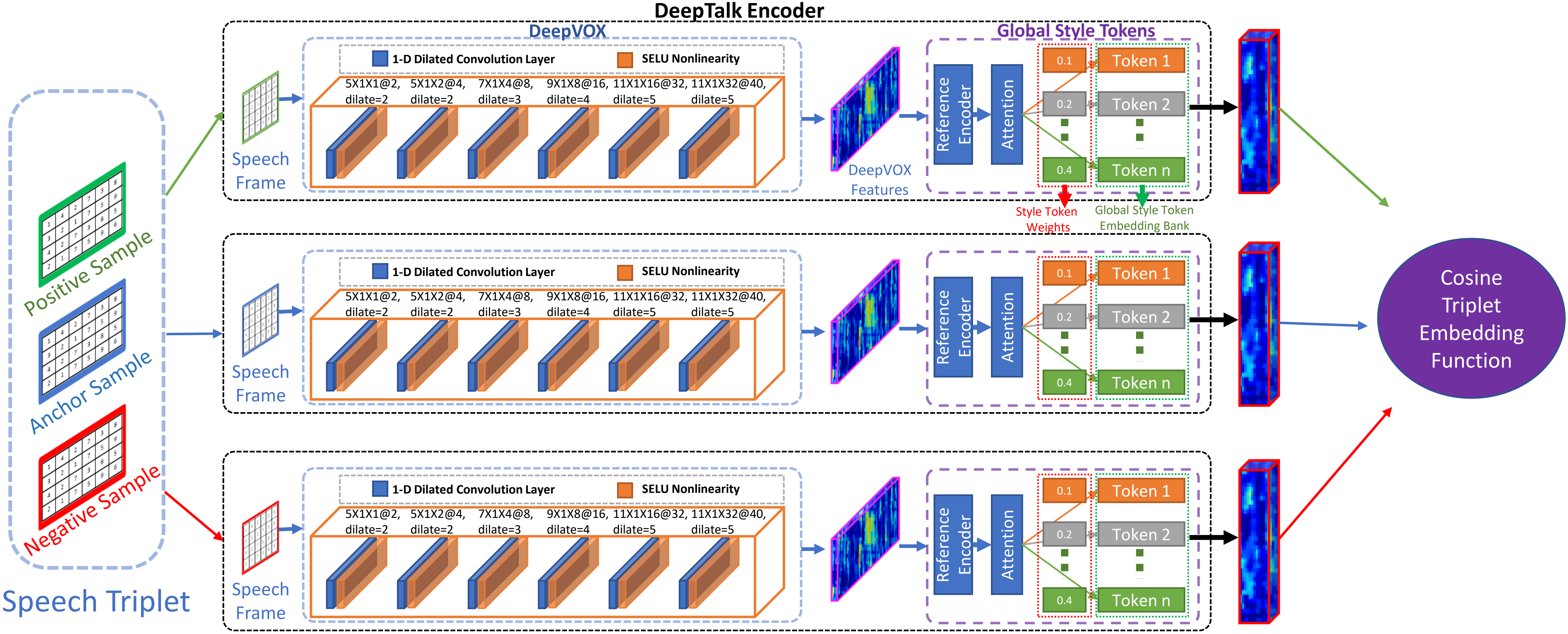
DeepTalk Encoder: Global Style Token (GST) based prosody embedding



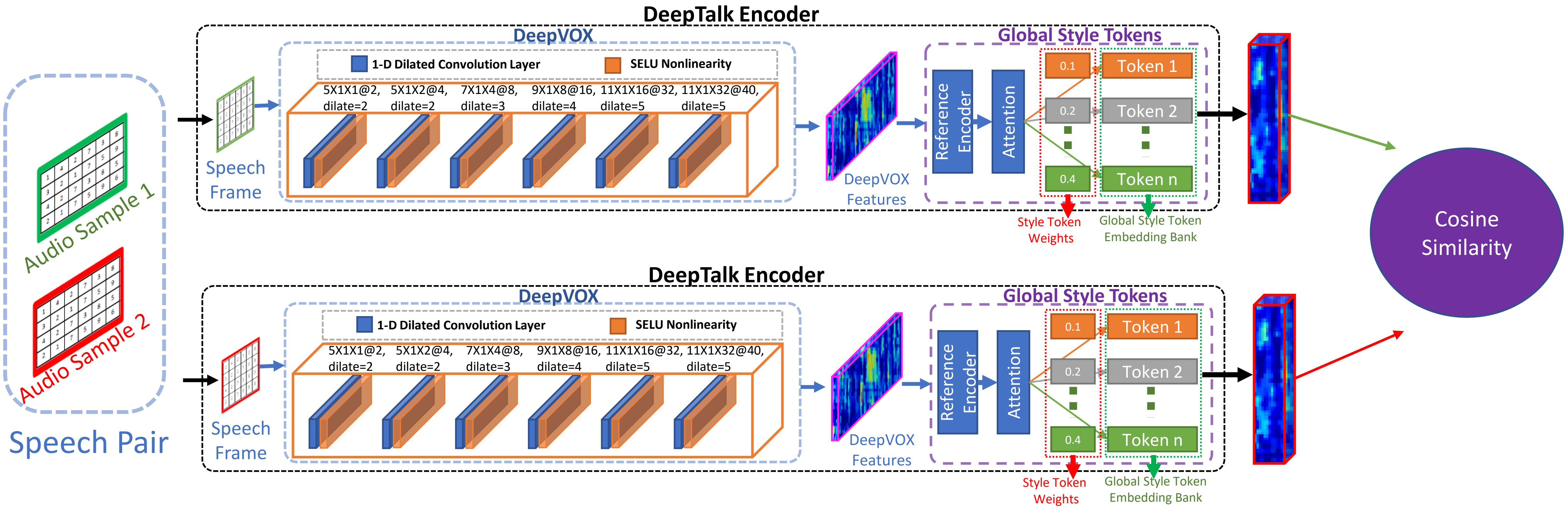
DeepTalk Encoder



DeepTalk Encoder – Training



DeepTalk Encoder – Testing



Datasets and Experiments

Datasets

VoxCeleb2 [1]

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

Type of Speech Data:
Interview Speech

NIST SRE 2008 [2]

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

Type of Speech Data:
Phone call and Interview Speech

NOISEX-92 [3]

Noise dataset:

Airplane (F16) Noise

Babble Noise

The average utterance length in both the VoxCeleb2 and NIST SRE 2008 datasets is around 5 secs

[1] Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." *arXiv preprint arXiv:1806.05622* (2018).
[2] "2008 NIST speaker recognition evaluation trainingset part 2 ldc2011s07,"https://catalog.ldc.upenn.edu/LDC2011S05, Accessed: 2018-03-06.
[3] 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.

Speaker Verification Experiments

Physiological Speech Feature-based Baseline Experiments

1) iVector-PLDA (MFCC)
2) xVector-PLDA (MFCC)
3) 1D-Triplet-CNN (MFCC-LPC)

Behavioral Speech Feature-based Experiments

4) The proposed DeepTalk method is used to perform vocal-style feature-based speaker verification experiments

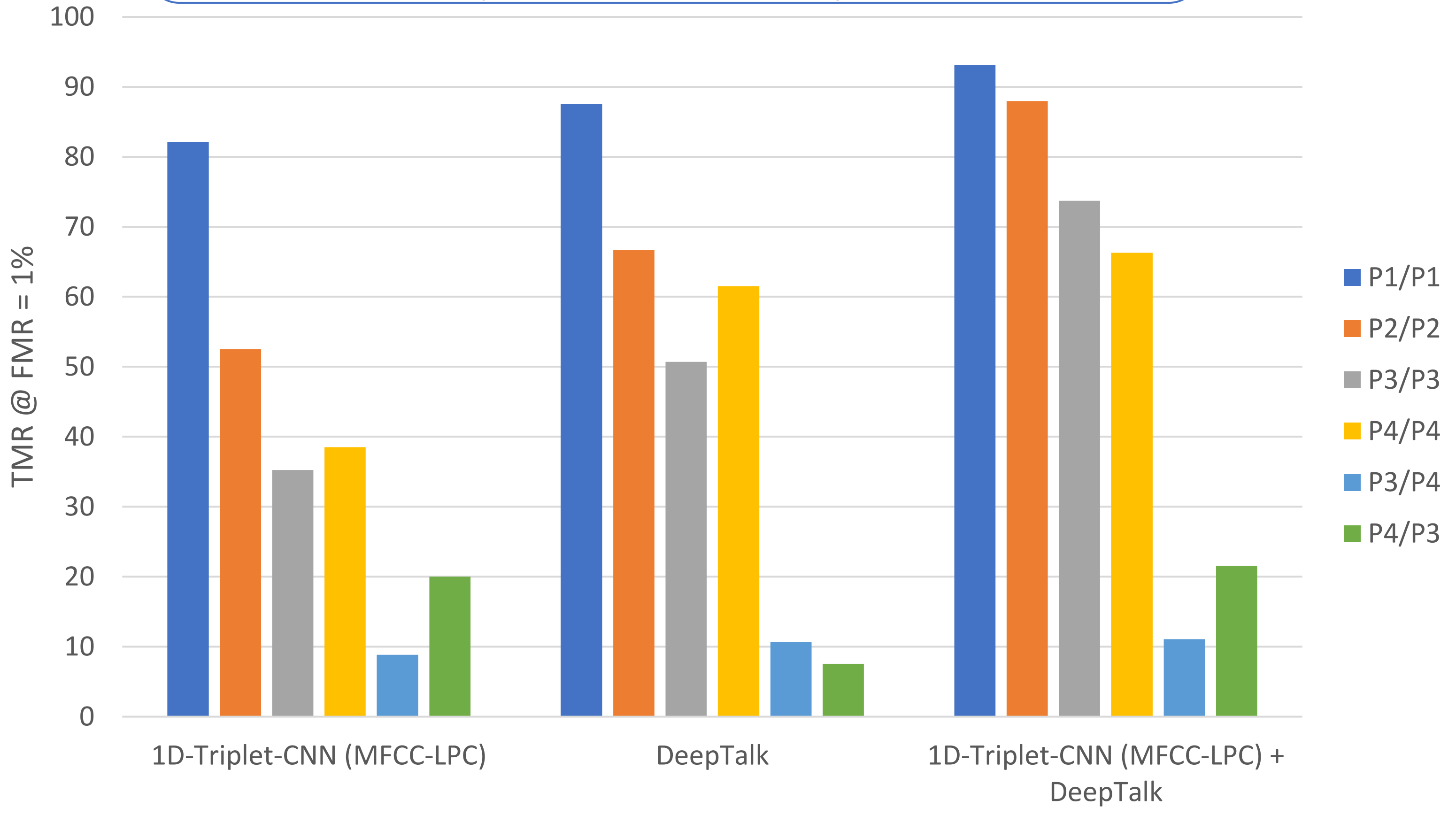
Combined physical and Behavioral Speech Feature-based Experiments

5) The DeepTalk and baseline methods are combined at a weighted score level, in a 1:3 ratio (chosen empirically), to evaluate the speaker recognition benefits of combining behavioral and physical speech features.

Speaker Verification Results

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%



Train / Test Data:

P1: VoxCeleb2

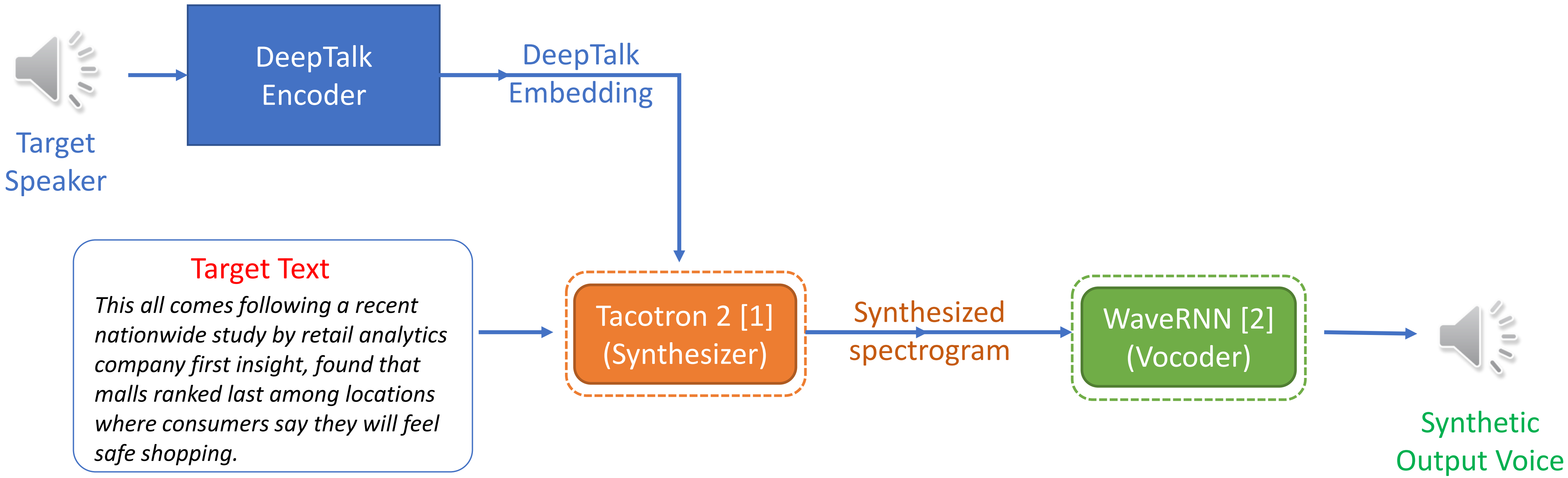
P2: NIST SRE 2008

P3: NIST SRE 2008 + Babble

P4: NIST SRE 2008 + F16

DeepTalk:
Vocal Style Encoding for Speech Synthesis

DeepTalk-based Speech Synthesis Framework



[1] Shen et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779-4783. IEEE, 2018.
[2] Kalchbrenner, et al. "Efficient Neural Audio Synthesis." In *International Conference on Machine Learning*, pp. 2410-2419. 2018.

Speech Synthesis Experiment

- We use DeepTalk to generate high-quality realistic synthetic speech using a target speaker's reference audio and a target text utterance
- We compare our results with synthetic speech generated using a **baseline Tacotron2** model

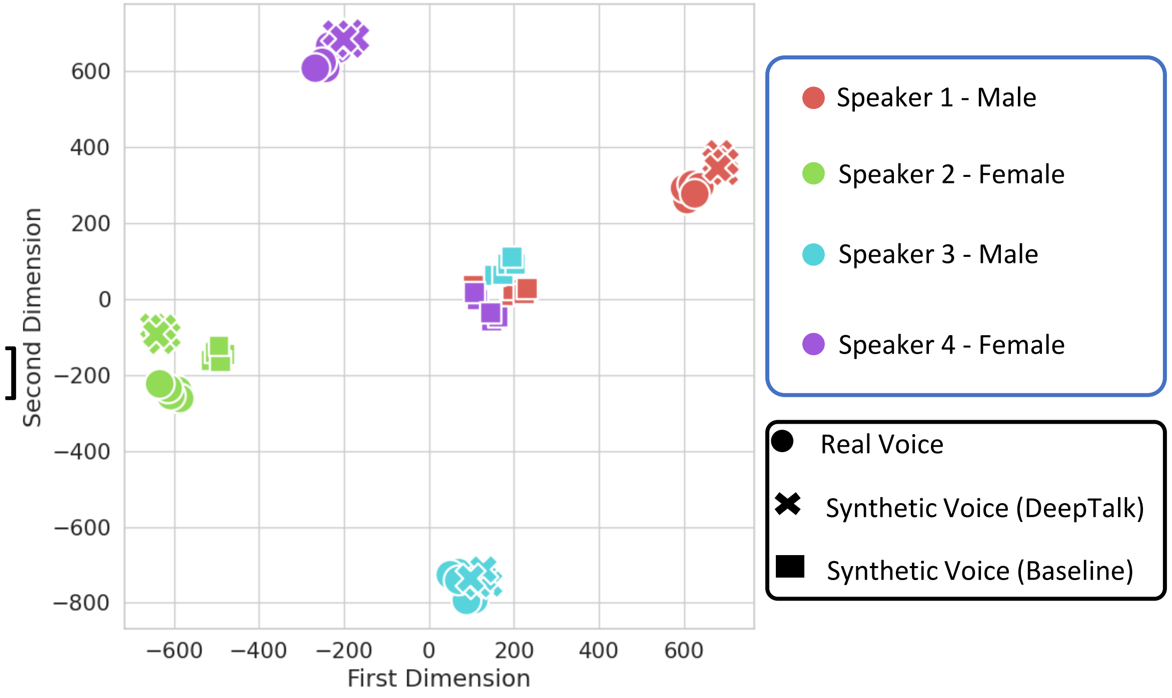
Target Text: In a scene that played out multiple times over the weekend and into Tuesday afternoon, the California National Guard airlifted hundreds of civilians

Target Speaker	**Reference Audio**	**Synthetic Audio (Baseline)**	**Synthetic Audio (DeepTalk)**
Speaker 1 Male			
Speaker 2 Female			

Note: The text utterances in the reference audios given above do not match the corresponding synthetic audios' utterances. The reference audios provide an example of the original voice of a given speaker. They can be used to compare the quality of vocal identity and style transfer in the corresponding synthetic audios.

t-SNE Plot-based analysis of DeepTalk

- 1D-Triplet-CNN-based speech embeddings are extracted from original and synthetic (both DeepTalk and baseline) speech samples for four different speakers.
- The speech embeddings are plotted in a t-SNE[1] plot
- DeepTalk-based synthetic speech samples are embedded closer to the Real Voice samples



[1] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. 11 (2008).

Possible Implication of Speech Synthesis

- Techniques like DeepTalk can improve the user-experience of Speech Generating Devices and digital voice assistants
- However, several concerns are raised by its potential misuse for creating DeepFake speech
- For example, in the past, DeepFake speech has been used to mimic an influential person's voice for defrauding[1]
- Therefore, such a technology should be used responsibly while adhering to appropriate privacy-protection laws

[1] Catherine Stupp, “Fraudsters used AI to mimic CEO’svoice in unusual cybercrime case,” The Wall Street Journal, vol. 30, 2019.

Summary

- Behavioral speech features extracted by DeepTalk method outperform majority of physical speech feature-based speaker verification methods
- Score-level fusion of DeepTalk with physical 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