DeepTalk: Vocal Style Encoding for Speaker Recognition and Speech Synthesis

Anurag Chowdhury\textsuperscript{1}, Arun Ross\textsuperscript{1}, Prabu David\textsuperscript{2}

\textsuperscript{1} - Department of Computer Science and Engineering
\textsuperscript{2} - College of Communication Arts and Sciences
Michigan State University
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]


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

Reference Audio

DeepTalk Encoder

DeepTalk embedding (Speaking Style)
DeepTalk Encoder Design

Reference Audio


DeepTalk Encoder Design

Reference Audio ➔ DeepVOX ➔ DeepVOX Features
DeepTalk Encoder Design: DeepVOX based speech feature extraction

Reference Audio

DeepVOX

DeepVOX Features

1-D Dilated Convolution Layer
SELU Nonlinearity
Pooling Layer
Loss Function
Alpha Dropout

5X1X1@2,
dilate=2
7X1X4@8,
dilate=3
11X1X16@32,
dilate=5

5X1X2@4,
dilate=2
9X1X8@16,
dilate=4
11X1X32@40,
dilate=5
DeepTalk Encoder: Global Style Token (GST) based prosody embedding
DeepTalk Encoder: 
Global Style Token (GST) based prosody embedding
DeepTalk Encoder

Input Audio

DeepTalk Embedding

Speech Frame

DeepVOX Features

Reference Encoder

Attention

Style Token Weights

Global Style Token Embedding Bank

DeepVOX

1-D Dilated Convolution Layer
SELU Nonlinearity
Pooling Layer
Loss Function
Alpha Dropout

5x1x1@2, dilate=2
5x1x2@4, dilate=2
7x1x4@8, dilate=3
9x1x8@16, dilate=4
11x1x16@32, dilate=5
11x1x32@40, dilate=5

Global Style Tokens

Token 1

Token 2

Token n
DeepTalk Encoder – Training

DeepTalk Encoder

DeepVOX

1-D Dilated Convolution Layer
SELU Nonlinearity

Speech Frame

5X1X1@2, 5X1X2@4, 7X1X4@8, 9X1X8@16, 11X1X16@32, 11X1X32@40

dilate=2, dilate=2, dilate=3, dilate=4, dilate=5, dilate=5

DeepVOX Features

Reference Encoder

Attention

Global Style Tokens

Token 1

Token 2

Token n

Style Token Weights

Global Style Token Embedding Bank

Speech Frame

5X1X1@2, 5X1X2@4, 7X1X4@8, 9X1X8@16, 11X1X16@32, 11X1X32@40

dilate=2, dilate=2, dilate=3, dilate=4, dilate=5, dilate=5

DeepVOX

1-D Dilated Convolution Layer
SELU Nonlinearity

Speech Frame

5X1X1@2, 5X1X2@4, 7X1X4@8, 9X1X8@16, 11X1X16@32, 11X1X32@40

dilate=2, dilate=2, dilate=3, dilate=4, dilate=5, dilate=5

DeepVOX

1-D Dilated Convolution Layer
SELU Nonlinearity

Speech Frame

5X1X1@2, 5X1X2@4, 7X1X4@8, 9X1X8@16, 11X1X16@32, 11X1X32@40

dilate=2, dilate=2, dilate=3, dilate=4, dilate=5, dilate=5

Cosine Triplet Embedding Function

Speech Triplet

Positive sample

Anchor sample

Negative sample
DeepTalk Encoder – Testing

DeepVOX

1-D Dilated Convolution Layer
SELU Nonlinearity

DeepTalk Encoder

Speech Frame

Reference Encoder
Attention

DeepVOX Features

Global Style Tokens

Embedding Bank

Style Token Weights

Global Style Token Embedding Bank

Cosine Similarity

Speech Pair

Audio Sample 1

Audio Sample 2
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

---

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.
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

This all comes following a recent nationwide study by retail analytics company First Insight, found that malls ranked last among locations where consumers say they will feel safe shopping.


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

<table>
<thead>
<tr>
<th>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</th>
<th>Synthetic Audio (Baseline)</th>
<th>Synthetic Audio (DeepTalk)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Speaker</strong></td>
<td><strong>Reference Audio</strong></td>
<td><strong>Speaker 1 Male</strong></td>
</tr>
</tbody>
</table>

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

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

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.