Effect of Speaking Style Variability on Speaker Verification

Nandita Raj Kumar, Cierra Yu

Background

- **Speaker verification** is identifying whether two different speech audio files are from the same speaker or not.
- The **ECAPA-TDNN** model is a pre-trained model for speaker verification trained on the VoxCeleb database using TDNNs and SE blocks.
- Speaker verification models face challenges with different **speaking styles** such as reading instructions, speaking in sentences, or making a phone call.
- **Equal error rate (EER)** is calculated as the point in which the false reject rate and the false acceptance rate are equal, and is used to evaluate speaker verification systems. A lower EER means the model is better at speaker verification.
- In this experiment, we test a speaker verification model's equal error rate (EER) when dealing with different speaking styles (reading instructions, speaking sentences, and making a phone call).

Research Question

How does speaking style variability affect the EER in speaker verification?

Hypothesis

- We predict that **variability in speaking style will reduce the accuracy of speaker verification** due to the change in pitch, speaking rate, and articulation when a person speaks in different manners.
- These **fluctuations result in a less accurate classification** of the speaker as the model may not recognize the variations to be from the same speaker due to its **different prosodic features**.

Methods

- Used the ECAPA-TDNN model created by Speechbrain to test our dataset.
- Dataset used was the UCLA Speaker Variability Database and was created by the UCLA Speech Processing and Auditory Perception Laboratory.
- Different tests included testing whether using different speech styles affected EER and whether the gender of speakers affected EER.
- Each test was run twice with 10 different speakers for each test.

Results

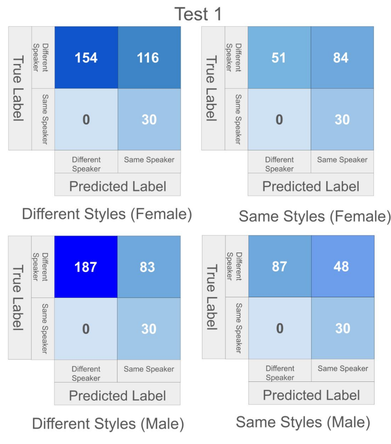

Fig. 1

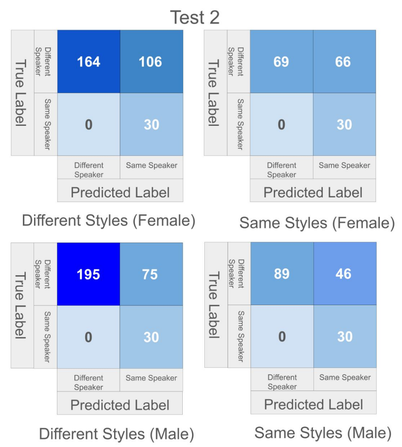

Fig. 2

Table 1: Evaluation metrics for females (Test 1)

Metric	F1	Accuracy	Precision	Recall	EER
Same Styles	0.416	0.490	0.263	1.0	0.383
Different Styles	0.340	0.613	0.205	1.0	0.300

Table 2: Evaluation metrics for females (Test 2)

Metric	F1	Accuracy	Precision	Recall	EER
Same Styles	0.476	0.6	0.312	1.0	0.328
Different Styles	0.361	0.646	0.220	1.0	0.281

Table 3: Evaluation metrics for males (Test 1)

Metric	F1	Accuracy	Precision	Recall	EER
Same Styles	0.555	0.709	0.384	1.0	0.262
Different Styles	0.419	0.723	0.265	1.0	0.235

Table 4 Evaluation metrics for males (Test 2)

Metric	F1	Accuracy	Precision	Recall	EER
Same Styles	0.6	0.757	0.428	1.0	0.228
Different Styles	0.508	0.806	0.340	1.0	0.176

Discussion

- **Tables 1 to 4** show that different speaking styles result in a lower EER. This may be because we tested with a small sample set and thus were unable to get optimal results. In addition, the EER score may be artificially low for different speaking style due to the inclusion of two similar speaking styles which increased the sample size.
- **Figures 1 & 2** are confusion matrices of different test sets and trials, all of which indicate a higher false positive rate in different styled trials.
- **Tables 1 to 4** show the significant difference in accuracy between male and female audio which is likely the result of an optimized pitch and formant analysis for male voices.
- **Tables 1 to 4** show that the model has 100% recall, meaning that the model accurately predicted all true positives samples.

Applications

- Speaker verification can be used as biometric security systems to verify or authenticate a user based on solely their voice. This form of security is used in many industries including law, medicine, and education.
- Speaker verification can be used to identify suspects for a criminal case with just a sample of an individual's voice at the time of crime.
- The medical industry uses speaker verification to reduce insurance fraud and cases by replacing current authentication methods.
- Speaker verification is used in education to help reduce cheating in standardized assessments by verifying students with voice.

Future Exploration

- How can speaker verification models be improved to differentiate speakers when whispering and/or with background noise?
- How can changes to the voice from aging be accounted for in speaker verification systems?

References

- Singh, Nilu & Khan, Prof. Raees & Pandey, Raj Shree. (2012). Applications of Speaker Recognition. Procedia Engineering. 38. 3122-3126. 10.1016/j.proeng.2012.06.363.
- https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/ecapa_tdnn