Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Sarina Meyer
Pascal Tilli
Florian Lux
Pavel Denisov
Julia Koch
Ngoc Thang Vu

Team IMS at VPC'22: A Cascade of Phonetic Speech Recognition, Speaker Embeddings GAN & Multispeaker Speech Synthesis

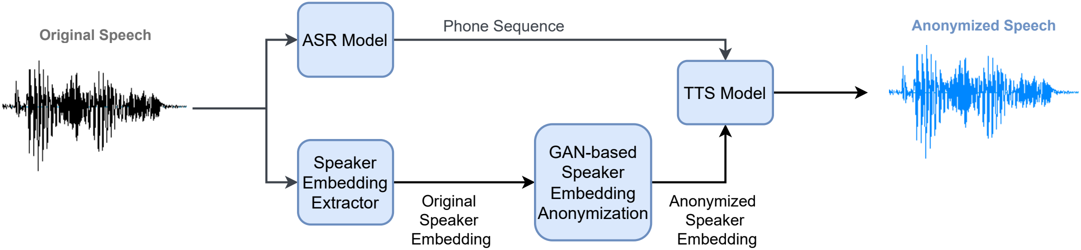
Our Idea

- Main problems of challenge baselines:
  - B1.a and B1.b: Usage of pitch and BN features $\rightarrow$ identity leakage
  - B2: Simple signal processing $\rightarrow$ not robust against neural attackers

- Our approach: Based on B1 pipeline but
  - **Phonetic Speech Recognition**
    - Reduction of speech to linguistic content; designed for optimal interaction with TTS
  - **Speaker Embedding Anonymization via GAN**
    - Generates artificial yet natural-like voices
  - **Multispeaker Speech Synthesis**
    - Optimized to produce distinctive voices based on speaker embedding

  $\rightarrow$ No usage of original pitch but instead smart pitch estimation

Speaker Anonymization Pipeline



Components: Speech Recognition

- Hybrid CTC/attention architecture [1] with Conformer encoder and Transformer decoder
- Implemented in ESPnet2 toolkit [2]
- Output: **phone sequences**
- Training transcriptions phonemized by IMS Toucan toolkit [3]
- Trained on LibriTTS [4]
  $\rightarrow$ used to label VoxCeleb corpora [5]
  $\rightarrow$ finetuned on VoxCeleb + LibriTTS
  $\rightarrow$ repeated 2x

Components: GAN Speaker Anonymization

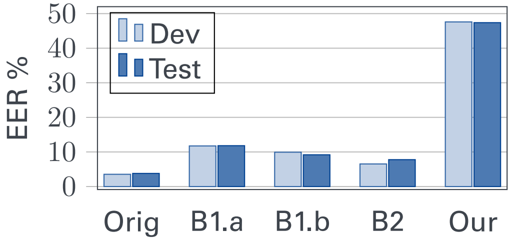
- Embeddings: Concatenation of **x-vector** [6] and **ECAPA-TDNN** [7] (704 dimensions)
  $\rightarrow$ extracted with SpeechBrain [8]
- **Wasserstein Generative Adversarial Network with Quadratic Transport Cost** [9] to generate artificial embeddings
  - Generator: transforms noise into 704-dimensional vector
  - Critic: distinguishes between real and fake data *distributions*
- During training: utterance-level speaker embeddings
- During inference: **one embedding per speaker** (exception: training data for eval models)

Components: Speech Synthesis

- **FastSpeech2 synthesis** [10] (phones $\rightarrow$ spectrograms) + **HiFiGAN vocoder** [11] (spectrogram $\rightarrow$ waveforms)
- Implemented in **IMS Toucan toolkit**
- Conversion of phone input into articulatory features
- Pitch and energy estimators based on FastSpeech2 and FastPitch [12]
- Training conditioned on concatenated speaker embeddings to produce different voices

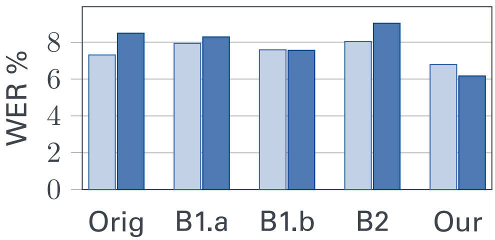
Results: Primary Evaluation

Privacy: ASV



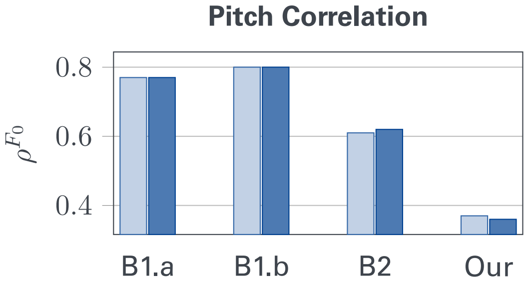
Regardless of the strong attacker: **almost perfect privacy**

Utility: ASR

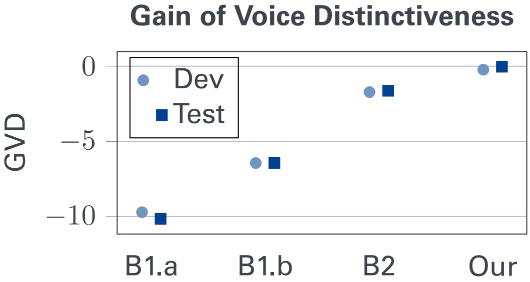


Best ASR results, even better than for original data
→ reduction of WER for VCTK from 12.82 to 7.81

Results: Secondary Evaluation



Low pitch correlation but $\rho^{F_0} > 0.3$ for all datasets



Distinctiveness of original data is **fully kept**

The Low Correlation of Pitch

- Our system **does not keep the original pitch sequences**
  $\rightarrow$ low pitch correlation scores
- This is deliberate:
  - Pitch contains too much speaker-identifiable information
  - Best for the system to have **no information** about the original prosody**about specific values of the original prosody**
- We actually do include prosodic information... in our **transcriptions**
  - ASR is trained on LibriTTS: outputs **punctuation**
  - The **context** and phonemized **word order** gives hints about intonation

  $\rightarrow$ The **energy and pitch estimation** based on that works pretty well!

Conclusion

- Our system: A speaker anonymization pipeline with ...
  - Phonetic ASR transcriptions
  - GAN-generated artificial but natural-like anonymous speaker embeddings
  - Multispeaker TTS with smart pitch estimation
- Highly outperform all baselines in 3 of 4 metrics:
  - **Almost perfect privacy** against strong attacker
  - **Better intelligibility** even than original VCTK data
  - **Same voice distinctiveness** as original data
- Deliberately without keeping pitch information to **reduce identity leakage**
  $\rightarrow$ nonetheless quite matching intonation

References I

[1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[2] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo et al., "The 2020 espnet update: new features, broadened applications, performance improvements, and future plans," in 2021 IEEE Data Science and Learning Workshop (DSLW). IEEE, 2021, pp. 1–6.

[3] F. Lux, J. Koch, A. Schweitzer, and N. T. Vu, "The IMS Toucan system for the Blizzard Challenge 2021," in Proc. Blizzard Challenge Workshop, vol. 2021. Speech Synthesis SIG, 2021.

[4] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in Interspeech, 2019, pp. 1526–1530.

[5] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," Computer Science and Language, 2019.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in ICASSP, 2018, pp. 5329–5333.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in Interspeech, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[8] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," 2021, arXiv:2106.04624.

References II

[9] H. Liu, X. Gu, and D. Samaras, "Wasserstein gan with quadratic transport cost," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4832–4841.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in International Conference on Learning Representations, 2020.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," NeurIPS, vol. 33, 2020.

[12] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in ICASSP. IEEE, 2021, pp. 6588–6592.