LPC based voice anonymisation

Skoltech

5 GOSLINGS

Our Plan

- Problem statement and motivation
- Speaker anonymisation algorithm using the McAdams coefficient
- NN-based method
- Comparing two approaches

Problem statement and motivation

Motivation

1. Privacy Protection, Confidentiality
2. Security and Anti-Fraud Measures

Problem statement

Obtaining anonymised user speech and comparing it with the original speech in terms of quality and anonymisation metrics

Algorithm Requirements

- **Content:** Focus on accurate speech content capture and reproduction, using Word Error Rate (WER) for assessment.
- **Sound Quality:** Quality of sound synthesis, typically evaluated through subjective mean opinion score (MOS) tests, is not a priority for measurement.
- **Voice Anonymity:** Ensuring the synthesized voice differs from the original to protect privacy, evaluated by cosine similarity.
- **Distinguish from Other Speakers:** The synthesized voice should be unique enough to be identifiable among others, can be checked through speaker verification

Metric

WER (Content quality)

$$WER = \frac{S + D + I}{N}$$

- S is the number of substitutions
- D is the number of deletions
- I is the number of insertions
- C is the number of correct words
- N is the number of words in the reference ($N = S + D + C$)

Cosine Similarity (difference between audios)

$$\frac{\langle u, v \rangle}{\|u\| \|v\|}.$$

LPC model

LPC - Linear Predictive Coding

LPC assumption

Let x_t be the amplitude of our signal at a given instant t.
According to the source-filter model, it's generated by a source signal e going through a resonant filter h:

$$x_t = (h * e)_t$$

LPC model assumes, that that the current signal depends on the past p samples and that the source is constant, so effectively:

$$x_t = \sum_{k=1}^{p} a_k x_{t-k} + e_t$$

LPC model

Encoding the whole signal

- Split signal into overlapping chunks.
- Assume that $x_i = 0$ for $i \leq 0$
- For each chunk solve linear system.
- Return matrices A and G, where $A[i,:]$ is solution for i-th chunk, and $G[i,:]$ is variance of the error e for it.

Decoding the whole signal

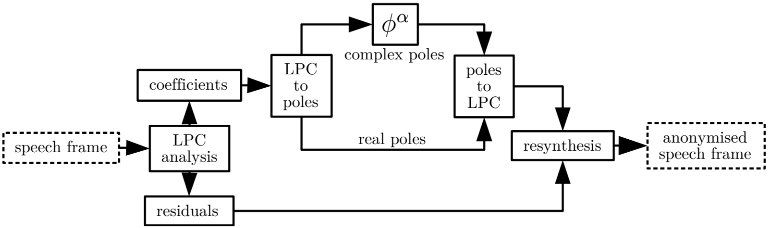
Decoding a LPC model consists in simulating a Source-Filter model: we first generate a source signal (white noise) and then apply a filter corresponding to the coefficients.

- Generate noise from normal distribution with variance G.
- Filter our signal A to initial chunks B
- Add chunks back and get initial signal x.

Related Works

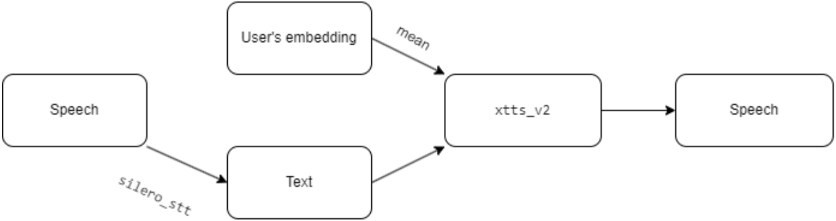
1. Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques [9]
    - utilize voice conversion techniques to anonymize speech data effectively
2. Towards directly modeling raw speech signal for speaker verification using CNNs [4]
    - model raw speech signals for speaker verification using Convolutional Neural Networks
3. Speaker anonimization using the McAdams coefficient [5]
    - paper about LPC implementation
4. Transfer learning from speaker verification to multispeaker text-to-speech synthesis [3]
    - transfer learning techniques from speaker verification to multispeaker text-to-speech synthesis

Speaker anonymisation algorithm using the McAdams coefficient



NN-based method

1. Two-step anonymization approach: speech-to-text and text-to-speech
2. We use Silero as speech-to-text model [8] and xTTSv2 as text-to-speech model [2]
3. User-embedding: mean of pre-trained embeddings, generate embeddings from normal distribution, use clean embeddings of some speaker and linearly combine embeddings



Original and anonymised text

Original text:
'Building a wall on the U.S.-Mexico border will take literally years'

original audio

after LPC anonymization

after NN-based method anonymization

More examples

Experiment

Validation Scheme

- Choose 100 texts from CommonVoice dataset [1]
- Use models to anonymize audio files
- Reconstruct text from these files using whisper-large-v3 [7]
- Compare texts in terms of WER
- Use convgru embedder [6]
- Compare audios in terms of cosine similarity of embeddings

- https://huggingface.co/microsoft/speecht5_tts
- https://huggingface.co/spaces/openai/whisper

Comparing two approaches

	Anonym. perf. Cosine sim. *smaller is better*	Text recon. perf. WER *smaller is better*	Time	Mem.
LPC	**0.734**	**0.198**	388ms	144MiB
NN: no anon.	0.685	0.277	1.53s	2126MiB (GPU)
NN: mean emb.	0.464	0.270		
NN: noise emb.	0.628	0.256		
NN: normal emb.	0.359	0.272		
NN: rand. speaker emb.	**0.345**	**0.279**		
NN: lin.comb 0.3	0.647	0.271		
NN: lin.comb 0.5	0.616	0.265		
NN: lin.comb 0.7	0.547	0.266		

Conclusions

- LPC works much faster than the NN-based method.
- LPC anonymization is the worst.
- When using random speaker embeddings, anonymization is better than in other approaches, but WER is at the same level with other NN-based methods.

Contribution of each team member

- Arkadiy: Reviewing literature on LPC-based methods for voice anonymization, implement LPC-based aprroach, verifying the algorithm's ability to anonymize voice (Wav2Vec embendings, cosine distance)
- Vsevolod: Reviewing literature, implementing Speaker anonimization algorithm using the McAdams coefficient, quality compare using WER metric
- Andrei: Implementation of NN-based voice anonymization method based, quality comparison
- Yuriy: Reviewing literature on NN-based methods for voice anonymization, presentation making, help with quality comparison
- Mikhail: suggested project idea, study what the architecture typical for ASR and TTS looks like, presentation making

¿PREGUNTAS?

References I

[1] Rosana Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*. 2020. arXiv: 1912.06670 [cs.CL].

[2] Gölge Eren and The Coqui TTS Team. *Coqui TTS*. Version 1.4. Jan. 2021. DOI: 10.5281/zenodo.6334862. URL: https://github.com/coqui-ai/TTS.

[3] Ye Jia et al. *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis*. 2019. arXiv: 1806.04558 [cs.CL].

[4] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcell. "Towards Directly Modeling Raw Speech Signal for Speaker Verification Using CNNS". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4884–4888. DOI: 10.1109/ICASSP.2018.8462165.

References II

[5] Jose Patino et al. *Speaker anonymisation using the McAdams coefficient*. 2021. arXiv: 2011.01130 [eess.AS].

[6] Christian Payer et al. *Instance Segmentation and Tracking with Cosine Embeddings and Recurrent Hourglass Networks*. 2018. arXiv: 1806.02070 [cs.CV].

[7] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. DOI: 10.48550/ARXIV.2212.04356. URL: https://arxiv.org/abs/2212.04356.

[8] Silero Team. *Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks*. https://github.com/snakers4/silero-models. 2021.

References III

[9] In-Chul Yoo et al. "Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques". In: *IEEE Access* 8 (2020), pp. 198637–198645. DOI: 10.1109/ACCESS.2020.3035416.