Speaker Anonymization Using X-vector and Neural Waveform Models

Fuming Fang[1], Xin Wang[1], Junichi Yamagishi[1] , Isao Echizen[1] , Massimiliano Todisco[2], Nicholas Evans[2], Jean-François Bonastre[3]
[1]National Institute of Informatics, Japan [2]EURECOM, France [3]University of Avignon, France

1. Background and aim

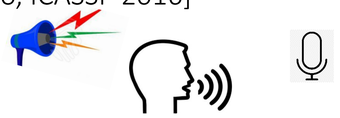
- Convenient web services allow us to share audio files
- Attackers may be able to synthesize high-quality voice of a user for spoofing purpose
- We propose a speaker anonymization method to conceal a speaker's identity before sharing audio files

Anonymization

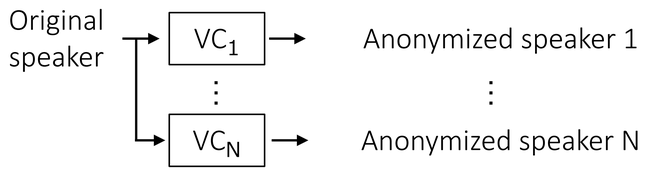
- Concealing speaker's identity
- Preserving linguistic content, quality, naturalness
- K-anonymization: same gender, age, etc → sounds natural
- Unable to be re-identified as the original speaker

2. Related work

- Adding background noise to distort speaker identity [Hashimoto, ICASSP 2016]



- Transforming a speaker into another speaker using voice conversion [Jin, ASRU 2009; Alegre, ICASSP 2014;Magarinos, Computer Speech & Language 2017]



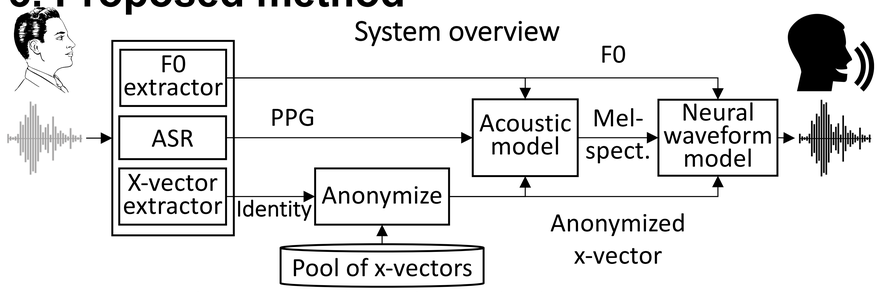
- ASR → speaker-dependent TTS [Justin, IEEE FG 2015]

Existing methods only focus on concealing speaker identity
Ours further consider content, quality, and naturalness

5. Conclusion & Future work

- Proposed to change x-vector for speaker anonymization
- Quality was preserved while identity was concealed
- WER and naturalness were not preserved when anonymized speaker largely differ from the original speaker
- Need to improve linguistic and speaker representation

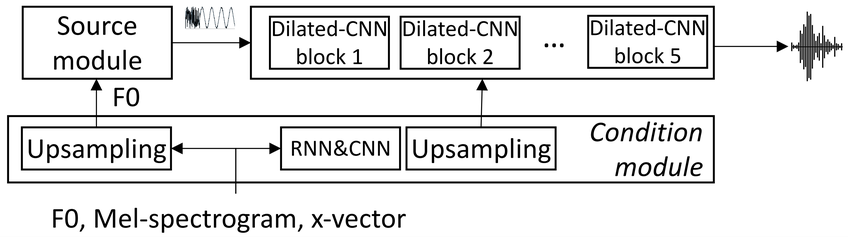
3. Proposed method



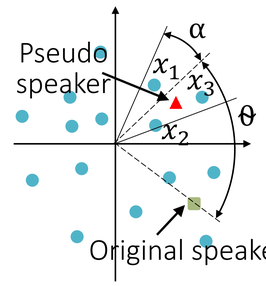
- Extract phoneme and speaker embeddings
- replace speaker embedding by averaged nearest K-speakers
- Generate high-quality waveform using neural vocoder

※ PPG: phoneme posteriorgram (ASR's softmax or hidden layer output, similar to phoneme embedding)

- Neural source filter [Wang, ICASSP 2019]



- X-vector anonymization



1. Dissimilarity score (distance): $1 - \cos(\theta)$
   If $\theta = 0$: nearest K-neighbors
2. Select a range of x-vectors
   $x_i \in range(\alpha)$
3. Generate a pseudo speaker:
   $\frac{1}{N} \sum_{x_i \in range(\alpha)} x_i$

PPG extraction	Kaldi TIMIT ASR (the 6th layer or softmax layer)
X-vector	6th layer (512 dimensions) X-vector pool: 7325 speakers
Acoustic model	FCs → LSTMs → FC → output (AR)

4. Experiment

Ideal anonymized speaker:

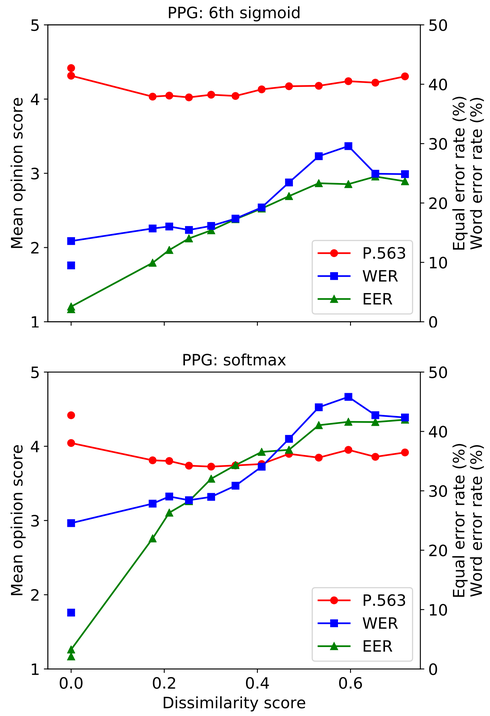
1. Should not be identified, ASV's EER↑
2. Quality (P.563 MOS) unchanged
3. WER (DeepSpeech) unchanged
4. Naturalness (MOS) unchanged

- EERs for K-neighbor anonymization (tested using an ASV system)

Nearest speakers for K-anonymization (voxceleb)	Nearest speakers for ASV comparison (VCTK)		
	3	6	9
-	2.52	2.04	2.04
100	23.00	20.77	19.74
200	24.81	21.92	21.66
300	25.05	22.89	22.60

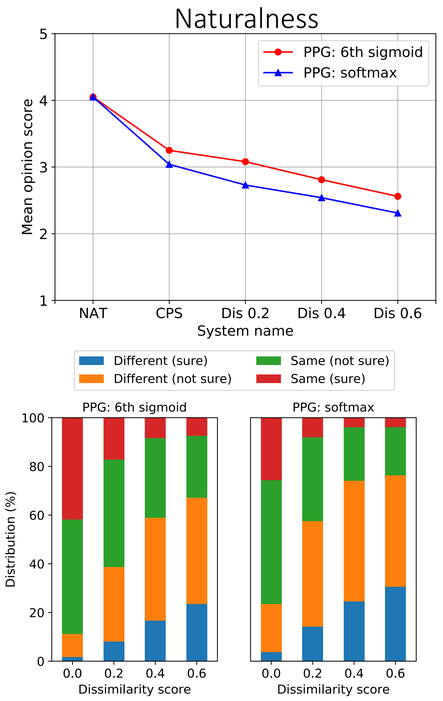
Anonymized speakers are different to the nearest neighbors to a certain degree

- Influence of different type of PPGs



Quality preserved while identity concealed

- Listening test result



Anonymized and the original one are different for human listeners