Use voice conversion for pseudonymisation?

Rob van Son

Netherlands Cancer Institute
ACLC, University of Amsterdam

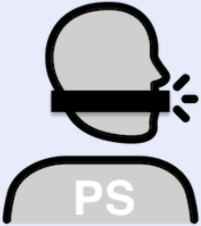
NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

UNIVERSITY OF AMSTERDAM
Amsterdam Center for Language and Communication

DELAD
28 January 2021

Sharing speech data

- Is needed for progress in speech science&technology [1]
- Privacy is a concern, especially for pathological speech
- Is *pseudonymization* possible?
  - Remove identity
  - *Retain linguistic & para-linguistic features*
  - What are the trade-offs?
- Applications:
  1. Demonstrations for live audience
  2. Speech corpora for study
  3. Fully Open Data
  4. *Secure processing in the cloud*



Introduction

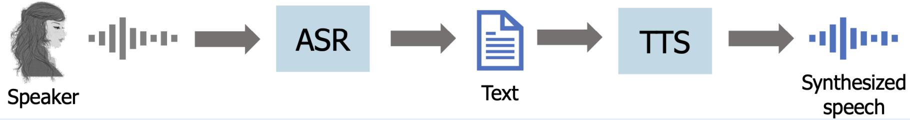
Pseudonymizing speech data

- *Anonymous*[1] means: *not identifiable by anyone*
- *Pseudonymous* means: *identifiable with extra information*
- The literature can be summarized as: [2, 3, 4]
  - *Anonymous data is not useful*
  - *Useful data is not anonymous*
- Before *Pseudonymous* speech can be shared, demonstrate:
  1. *Security*
  2. *Usefulness*

[1] There is a lot of legal ambiguity and uncertainty here, see [4]

Approaches to Pseudonymizing speech data

ASR⇒TTS



- Remove speaker identity? yes
- Keep unchanged other characteristics (e.g., prosody, emotions)? no
- Preserve linguistic content? yes, but not perfectly due to ASR (acoustic model) errors
- Diversity and distinguishability of synthesized voices? limited

N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Approaches to Pseudonymizing speech data

Signal processing

- Remove speaker identity? less well
- Keep unchanged other characteristics (e.g., prosody, emotions)? yes
- Preserve linguistic content? yes
- Diversity and distinguishability of synthesized voices? limited

Adapted from N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Speaker Identity

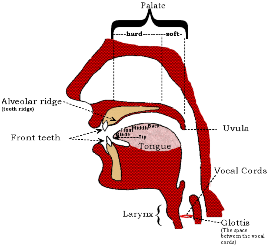
Partition identity in speech

- *Inherent* features
  derived from a speaker's anatomy and physiology (vocal tract length)
- *Learned* features
  acquired during language learning and use (dialect, accent)
- *Linguistic* features
  depend on the message and pragmatics (register)

Pseudonymization targets

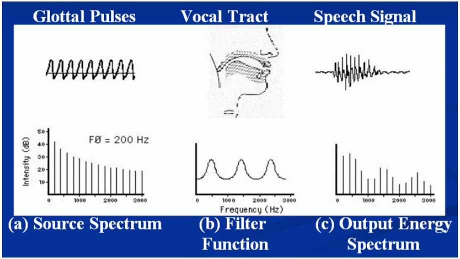
- *inherent* features
- ~~*learned* features~~

Speaker Identity: Anatomy to Sound

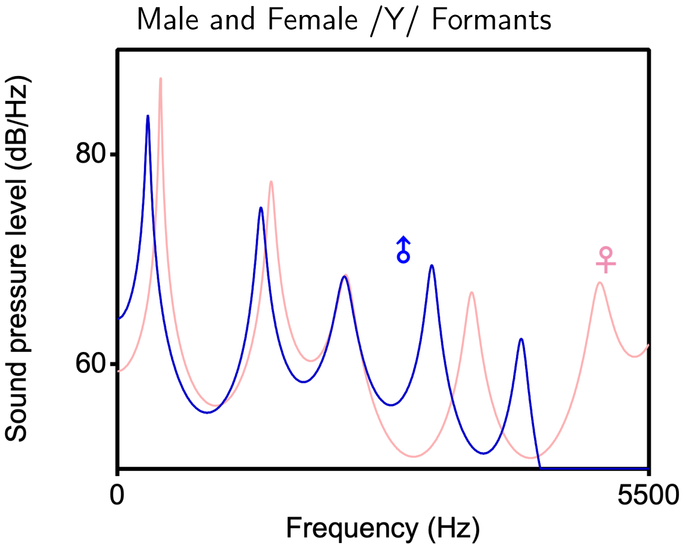


$\Rightarrow$

Vocal Tract Tube Model

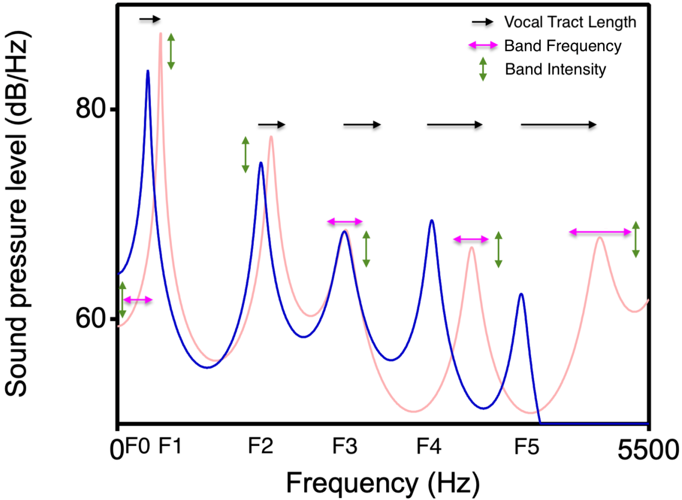


Example Pseudonymization: Signal processing



Example Pseudonymization: Signal processing

Change from Blue to Pink (or vv.)



Signal Processing: Vocal Tract Length and sound changes

- Estimate source Vocal Tract length (VTL) from formants [5]
  (requires ∼300 seconds of speech)
- Change recording to target VTL, pitch, speed [6]:
  `Change gender: 75, 600, 1.2, 120, 1, 0.9` (Praat)
- Shift bands F_0, F_3-F_5 to target frequencies
- (De-)amplify bands F_0-F_5 to target intensities

⇒ *create sound with desired F_x formant frequency and intensity*

⇒ *splice F_x band into target sound*

https://robvanson.github.io/PseudonymizeSpeech

Example Experiment: Online, self paced

The voices of the speakers A and B have been changed.
Which one do you think is the unknown speaker X?

Long VT	A	X	B
Short VT	A	X	B

https://robvanson.github.io/akouste/ABX_ListeningExp.html [7]

Example Experiment: Online, self paced

The voices of the speakers A and B have been changed.
Which one do you think is the unknown speaker X?

Long VT	A	X	B
Short VT	A	X	B

https://robvanson.github.io/akouste/ABX_ListeningExp.html [7]

Example Experiment: Online, self paced

The voices of the speakers A and B have been changed.
Which one do you think is the unknown speaker X?

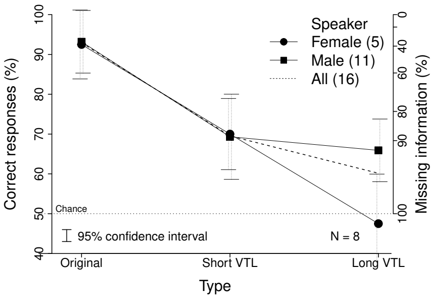
Long VT	A	X	B
Short VT	A	X	B

https://robvanson.github.io/akouste/ABX_ListeningExp.html [7]

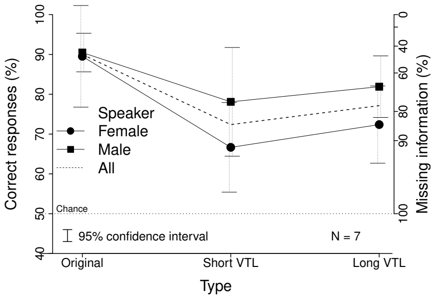
Example Experiment: Results

ABX listening experiments

⇐ Pseudonymized | De-pseudonymized ⇒



Responses to pseudonymized speech



Responses after →de-pseudonymization
15F/15M speakers for each Type, 90 in total

Original: AB are original recordings, *Short VTL*: AB pseudonymized to a short vocal tract length, *Long VTL*: AB pseudonymized to a long vocal tract length.

Dubagunta et al. (submitted) [8]

Summary 1: Example pseudonymization

Results for human listeners

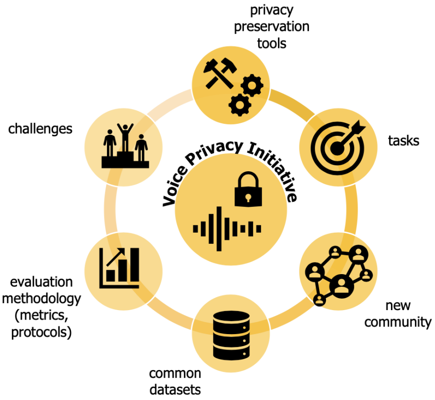
- >80% of identifying information can be removed from speech
- Speech quality is good to near natural

But many questions remain

- Is Pseudonymization reversible?
- Is *Automatic Speaker Identification* still possible?
- Are para-linguistic features preserved?
- Are speech pathologies preserved?

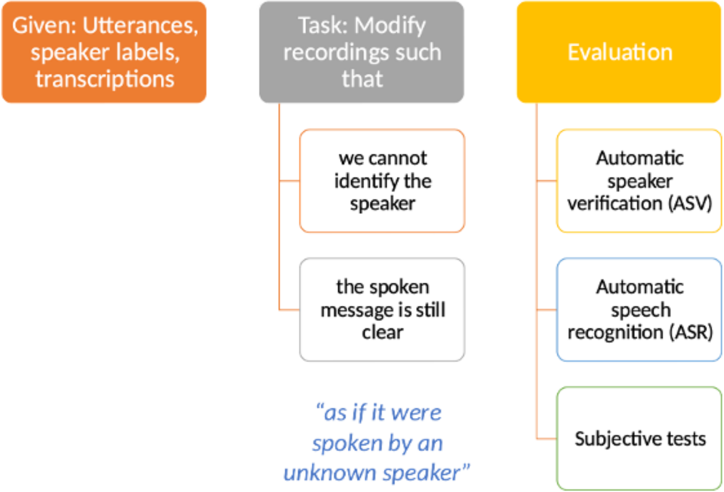
Voice Privacy Challenge 2020: Aim

Promote the development of privacy preservation tools for speech technology



N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Voice Privacy Challenge 2020: Task



Slide courtesy S. Pavankumar Dubagunta

Voice Privacy Challenge 2020

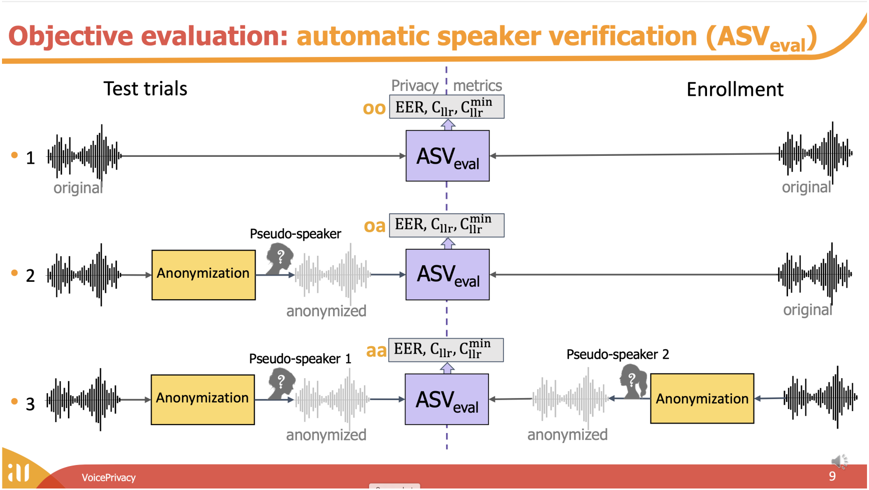
Evaluation metrics [9]

- Privacy: Automatic Speaker Verification (ASV_{eval})
  Equal Error Rate - EER $= P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER})$
- Utility: Automatic Speech Recognition (ASR_{eval})
  Word error rate - WER $= 100 \cdot \frac{N_{sub}+N_{del}+N_{ins}}{N_{ref}}$
- Subjective listening tests
- Utility: ~~Para-linguistic speech classification~~ (Not yet)

fa: false alarm rate

N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Voice Privacy Challenge 2020

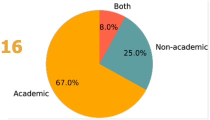


N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Voice Privacy Challenge 2020

Participants

- Registered teams: **25** (more than **45** participants) from **13** countries
- Teams submitted valid results: **7** (**+1** contribution related to evaluation models)
  - deadline-1: submissions from **6** teams
  - deadline-2: submissions from **3** teams
- Submitted anonymization systems: **16**



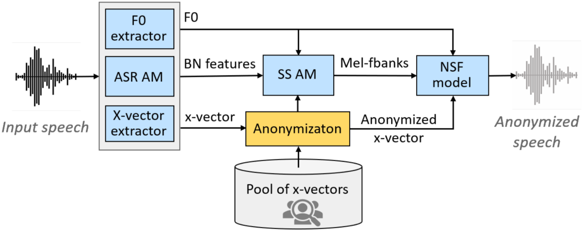
- Post-evaluation analysis (submission of the anonymized dataset for training evaluation models): **4**

Team	Country	Status
Idiap-NKI	Switzerland	academic
Biometric Vox	Spain	non-academic
DA-IICT Speech Group	India	academic
Team SDU	Turkey	academic
PingAn	USA	non-academic
AIS-lab JAIST	Japan / Thailand	academic
BlackBox@CMU	USA	academic
Motorola Solutions	USA	non-academic
MultiSpeech	France	academic
Orange ITAAC Team	France	non-academic
Oxford System Security Lab	UK	academic
Preech	USA	academic
Sigma Technologies S.L.U.	Spain	both
TMU	Japan	academic
Ioenix	USA	non-academic
VTouch	China	academic
VIAX	China	academic
PhoneClearly.com	USA	non-academic
Kyoto Team	Japan	academic
PSUT	Jordan	academic
TJU-VP	China	academic
EAM AAU ANONYMOUS	Denmark	academic
TalkMeUp	USA	both
Fearghal Sheehan	Ireland	academic

Late registration

N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Baseline 1 Anonymization using x-vectors and neural waveform models



- **ASR AM**: Automatic speech recognition acoustic model (to extract **BN** (bottle-neck) features)
- **SS AM**: Speech synthesis acoustic model
- **NSF**: Neural source-filter model

https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

Inspired from: [Fang 2019]

VoicePrivacy Screenshot 17

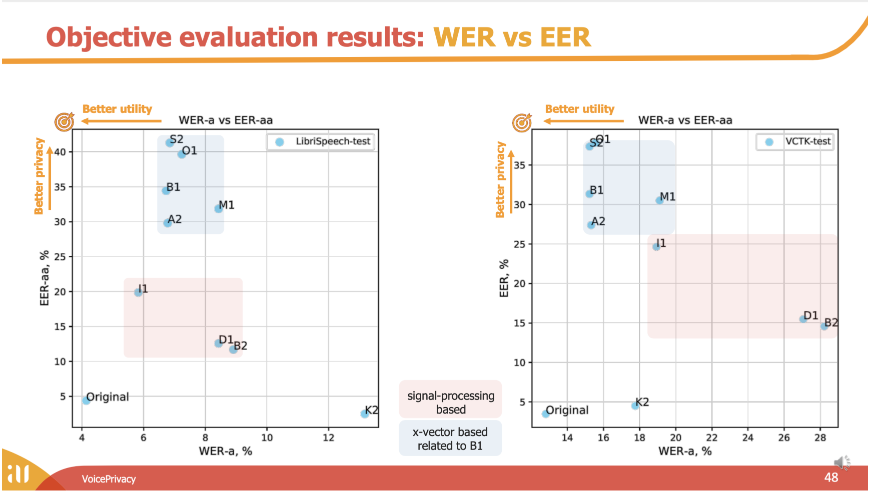
N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org

Voice Privacy Challenge: Baseline examples

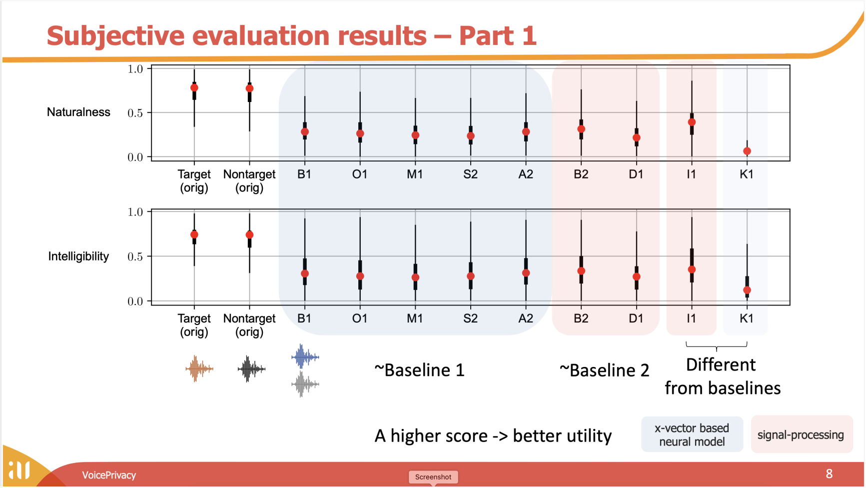
		Original	Anonymized
LibriSpeech	Female		
	Male		
VCTK	Female		
	Male		

URL: www.voiceprivacychallenge.org

Voice Privacy Challenge 2020

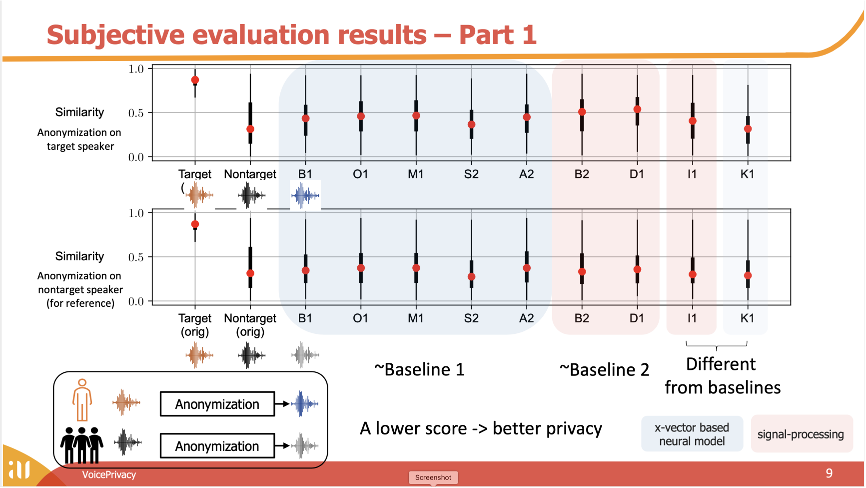


N. Tomashenko et al. 2020; URL: www.voiceprivacychallenge.org



Xin Wang et al. 2020; URL: www.voiceprivacychallenge.org

Voice Privacy Challenge 2020



Xin Wang et al. 2020; URL: www.voiceprivacychallenge.org

Summary 2: Voice Privacy Challenge 2020

Results for Automatic Speaker Verification

- Identifying information can be removed from speech (EER∼30%-40%)
- Intelligibility is reduced (WER∼6%-20%)
- Speech quality is (strongly) affected (subjective)

Questions remain

- Are results corpus- and style-dependent?
  (ASV might learn reading peculiarities of speakers?)
- Can para-linguistic features be preserved?
- Can speech pathologies be studied?
- What are the trade-offs?

Beyond the Challenge

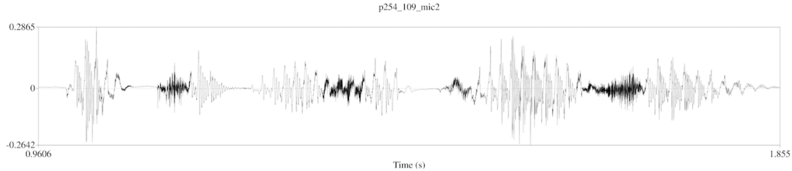
Make Voice Privacy useful for studying (pathological) speech

- Privacy can be protected, improvements are still welcome
- Can formants be measured in pseudonymized speech?
- Can pathological speech, e.g., dysarthric speech, still be studied?

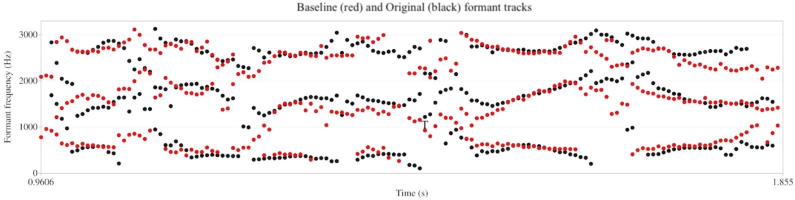
Dubagunta et al. (submitted) [8]

Formant tracks after pseudonymization

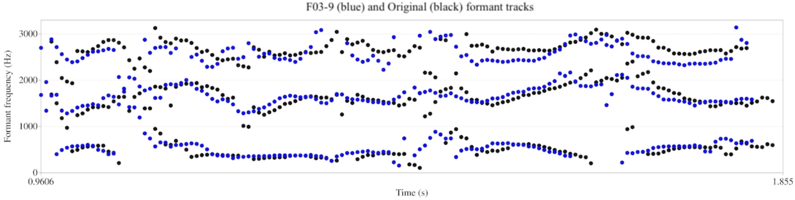
Original



Baseline



NKI-
Idiap



[9]

Formant tracks after pseudonymization

Formant track "movements" are preserved to some extent, but the pseudonymization method and speaker gender matter.

Mean correlation coeff, R (SD), between formant tracks from Original and pseudonymized recordings, for all speakers (N=30)

Group	Method	F_1	F_2	F_3
F (15#)	Baseline	0.507 (0.158)	0.601 (0.198)	0.424 (0.287)
	NKI-Idiap	0.563 (0.194)	0.659 (0.161)	0.620 (0.202)
M (15#)	Baseline	0.490 (0.161)	0.571 (0.158)	0.264 (0.226)
	NKI-Idiap	0.655 (0.153)	0.716 (0.136)	0.688 (0.136)
Total	Baseline	0.499 (0.160)	0.586 (0.178)	0.344 (0.257)
	NKI-Idiap	0.609 (0.174)	0.688 (0.149)	0.654 (0.169)

Dysarthria classification after pseudonymization

TORGO corpus: 8 Dysarthric + 7 Control [10]

- Pseudonymize recordings
- Train Dysarthria classifiers on Original and Pseudonymized speech
- Compare classification results on individual sentences

Data

- Recording quality low
- Start with 30 sessions and 15 speakers
- Classifier fails on 15 sessions ($<$70% correct on Original)
- Keep 15 sessions from 10 speakers

Dysarthria classification after pseudonymization: examples*

	Original	Pseudonymized
Female (F01)		
Male (M01)		

*NKI-Idiap pseudonymization

URL: https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html [10]

Dysarthria classification after pseudonymization

Results: % correct classification on the TORGO corpus

Group	Speaker	Original	Pseud.	Conc.	N
Control	FC01	98.2	47.6	49.4	164
	FC02	86.3	13.7	24.4	1000
	MC01	98.5	99.3	98.4	748
	MC02	99.1	98.7	98.3	464
	MC03	99.3	100.0	99.3	600
Dysarthric	F01	90.2	90.9	90.2	132
	M01	92.7	99.7	92.4	288
	M02	95.8	98.5	95.8	409
	M03	91.3	97.9	91.5	424
	M04	91.0	93.6	87.5	488
Total		94.2	84.0	82.7	471.7

Only sessions&speakers with $\geq$70% correct for Original. Conc.: Concordance, percentage identical classification. N: # utterances. Cronbach's α=0.769 (all), α=0.949 (excl. FC01&02)

Dysarthria classification is preserved for some speakers.

Summary 3: Beyond the Challenge

Results

- Formant tracks can be preserved to some extent
- Dysarthria classification can be preserved for some speakers

Problems

- Sensitivity to audio quality
- Speaker specific performance
- Are universal algorithms possible?

Dubagunta et al. (submitted) [8]

Conclusions

Results are encouraging

- Speaker identity can be hidden
- Intelligibility can be good
- Speech quality should be improved
- Para-linguistic aspects can be preserved, but need work

Next step: A case study (challenge?)

- A para-linguistic speech task
- Good privacy, intelligibility & quality (quantified)
- Good para-linguistic speech classification or grading

⇒ e.g., Emotion, PD, Dysarthria, ...

Thank You

?

This research was conducted together with

S. Pavankumar Dubagunta and Mathew Magimai-Doss from Idiap, Switzerland

DOI: 10.5281/zenodo.4452824

The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives an unrestricted research grant of Atos Medical AB, Hörby, Sweden

Literature

[1] Alison Ferguson, Hugh Craig, and Elizabeth Spencer. "Exploring the Potential for Corpus-Based Research in Speech-Language Pathology". en. In: *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*. 2009, pp. 30–36.

[2] Ira S Rubinstein and Woodrow Hartzog. "Anonymization and Risk". en. In: *WASHINGTON LAW REVIEW* 91 (2016), p. 59.

[3] Sophie Stalla-Bourdillon and Alison Knight. "Anonymous Data v. Personal Data – A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data". en. In: *Wisconsin International Law Journal* 34.2 (2017), p. 39.

[4] Michele Finck and Frank Pallas. "They who must not be identified—distinguishing personal from non-personal data under the GDPR". en. In: *International Data Privacy Law* 10.1 (2020), pp. 11–36.

Literature

[5] Adam C. Lammert and Shrikanth S. Narayanan. "On Short-Time Estimation of Vocal Tract Length from Formant Frequencies". In: *PLOS ONE* 10.7 (2015), e0132193. DOI: 10.1371/journal.pone.0132193.

[6] Paul Boersma and David Weenink. *Praat: a system for doing phonetics with the computer*. 2017. URL: http://www.praat.org.

[7] Xin Wang et al. "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech". In: *Computer Speech & Language* 64 (2020). Publisher: Elsevier, p. 101114. DOI: 10.1016/j.csl.2020.101114. URL: https://hal.archives-ouvertes.fr/hal-02945493.

[8] S. Pavankumar Dubagunta, Rob J.J.H. van Son, and Mathew Magimai-Doss. "Adjustable Deterministic Pseudonymization of Speech". In: (). Submitted.

Literature

[9] Natalia Tomashenko et al. "VoicePrivacy 2020 Challenge Evaluation Plan". en. In: *Odyssey 2020*. 2020, pp. 1–21. URL: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf.

[10] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. "The TORGO database of acoustic and articulatory speech from speakers with dysarthria". In: *Language Resources and Evaluation* 46.4 (2012), pp. 523–541.