Hiding speaker’s sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline

Paul-Gauthier Noé[1], Xiaoxiao Miao[2], Xin Wang[2], Junichi Yamagishi[2], Jean-François Bonastre[1] and Driss Matrouf[1]
[1]Laboratoire Informatique d’Avignon (LIA) – Avignon Université, France [2]National Institute of Informatics (NII), Japan

Introduction

Objective: Hide in speech utterances the information related to the sex of the speaker for **privacy** purposes,
Why: Speech technologies are spreading and privacy considerations are rising consequently,
What we propose: An analysis / synthesis voice conversion with an affine transformation of the f_0 and a zero-LLR transformation of the speaker representation.

Some prerequisites

Bayesian updating of belief:

Considering a set of classes $\mathcal{C} = \{c_0, c_1\}$ (let’s say male and female) and an attacker who wants to infer the class of an observation x, its posterior probabilities are given by:

$$\text{logit}\, P(c_i|x) = \log \frac{P(x|c_i)}{P(x|c_{\neg i})} + \text{logit}\, P(c_i),$$

where $i \in \{0, 1\}$ and $P(c_i)$ is the attacker’s prior.

Shannon’s perfect secrecy:

Posterior and prior probabilities must be equal [1]:

$$\forall\, i,\ P(c_i|x) = P(c_i)$$

→ Set the log-likelihood-ratio (LLR) to 0.

Neural discriminant analysis for zero-evidence

In [2], a discriminant analysis has been proposed for the control of the LLR. This has been used for the concealment of the information related to the sex of the speaker in speaker embeddings

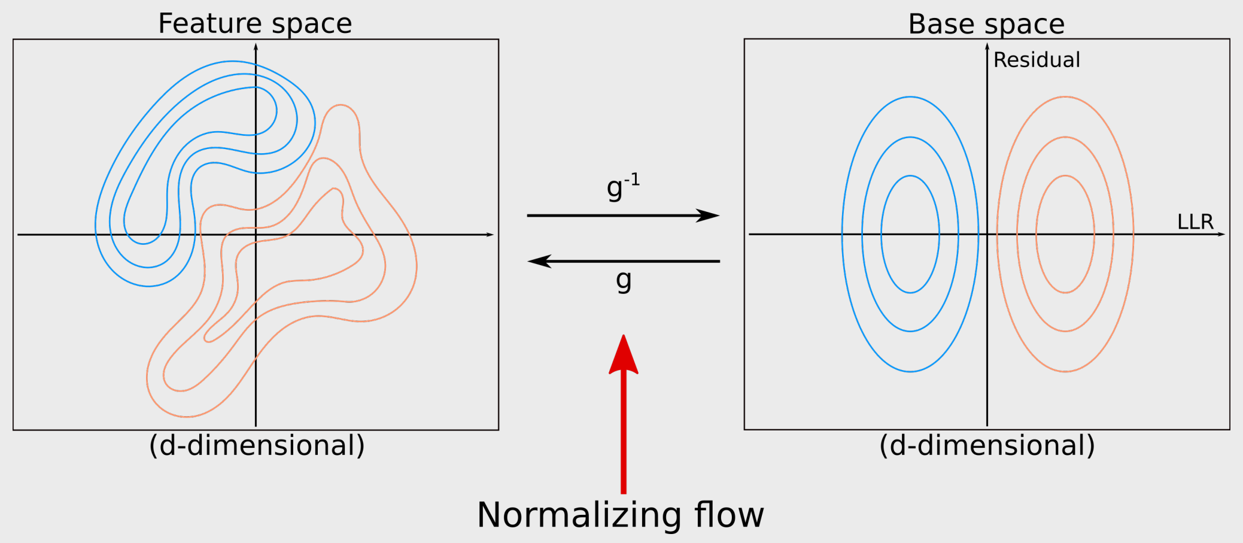


Figure: Illustration of the discriminant analysis in question.

In the base space, the LLR can be set to zero for privacy.

References

[1] *Claude E. Shannon.*
Communication theory of secrecy systems.
The Bell System Technical Journal, *28(4), 1949.*

[2] *Paul-Gauthier Noé et al.*
A bridge between features and evidence for binary attribute-driven perfect privacy.
In Proc. ICASSP, *pages 3094–3098, 2022.*

[3] *Natalia Tomashenko et al.*
Introducing the VoicePrivacy Initiative.
In Proc. Interspeech, *pages 1693–1697, 2020.*

[4] *Paul-Gauthier Noé et al.*
Towards a unified assessment framework of speech pseudonymisation.
Computer Speech & Language, *72:101299, 2022.*

Proposed protection system

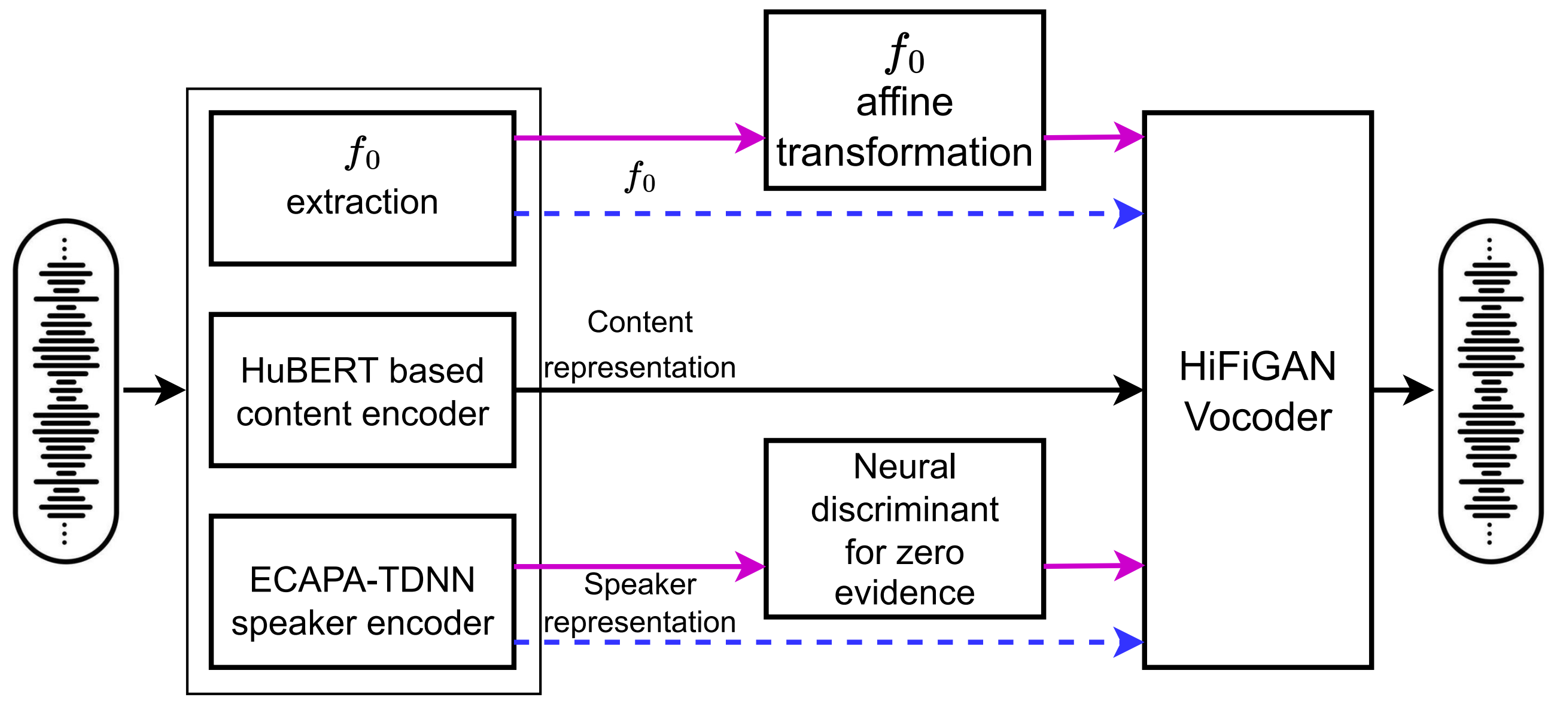

f_0 extraction
HuBERT based content encoder
ECAPA-TDNN speaker encoder
f_0
f_0 affine transformation
Content representation
Neural discriminant for zero evidence
Speaker representation
HiFiGAN Vocoder

Figure: Architecture of our system (blue-dashed path during training and purple during protection).

- **Speaker representation protection** is done by applying the neural discriminant analysis-based zero-evidence-based protection introduced in the prerequisites.
- **f_0 protection** is done by applying an affine transformation on the f_0 to impose a fixed f_0 mean and standard deviation for all utterances. These target f_0 moments are computed to be "sex-independent" i.e. "in between" males and females.
- **The content** is assumed to not contain sex-related information.

Experiments

Our system is compared with the following **baselines**:

- *TDPSOLA* is just an affine transformation of f_0 using the time domain pitch synchronous overlap add algorithm,
- *global* is the same as our analysis / synthesis, but instead of the neural discriminant-based protection, the speaker embeddings are replaced by a single averaging of speaker embeddings between males and females,
- *synthesized* is copy-synthesis.

Results:

Automatic sex classification for protection assessment

The used classifier is based on HuBERT base features extraction, statistical pooling, and multilayer perceptron.

Table: Sex classification results for protection assessment.

	ignorant		semi-informed	
system	EER [%] → 50%	D_{ECE} [bit] ↓	EER [%] → 50%	D_{ECE} [bit] ↓
original	3.67	0.578		
synthesised	4.32	0.542	4.01	0.593
global	24.95	0.198	20.60	0.233
TDPSOLA	6.30	0.504	4.36	0.542
proposed	28.99	0.128	24.13	0.200

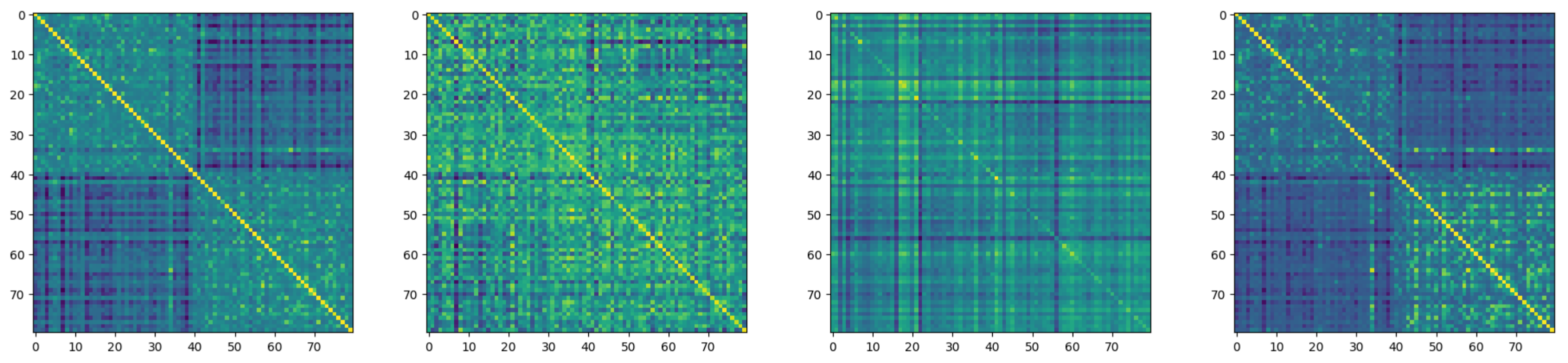
Automatic speech recognition (ASR) and speaker verification (ASV)

The used ASR and ASV systems are those from the VoicePrivacy Challenge evaluation plan [3].

Table: Automatic speech recognition and speaker verification results.

	ASR	ASV	
system	WER [%] ↓	EER [%] ↓	$C_{\text{llr}}^{\text{min}}$ [bit] ↓
original	4.02	5.77	0.204
synthesised	4.79	6.86	0.240
global	4.92	35.86	0.903
TDPSOLA	4.43	6.38	0.237
proposed	4.81	11.55	0.407

Voice similarity matrices



(a) original (b) proposed (c) global (d) TDPSOLA

Figure: Voice log-similarity matrices [4].

Listening tests

For both **naturalness** and **protection** assessment (19 listeners, all native English speakers).

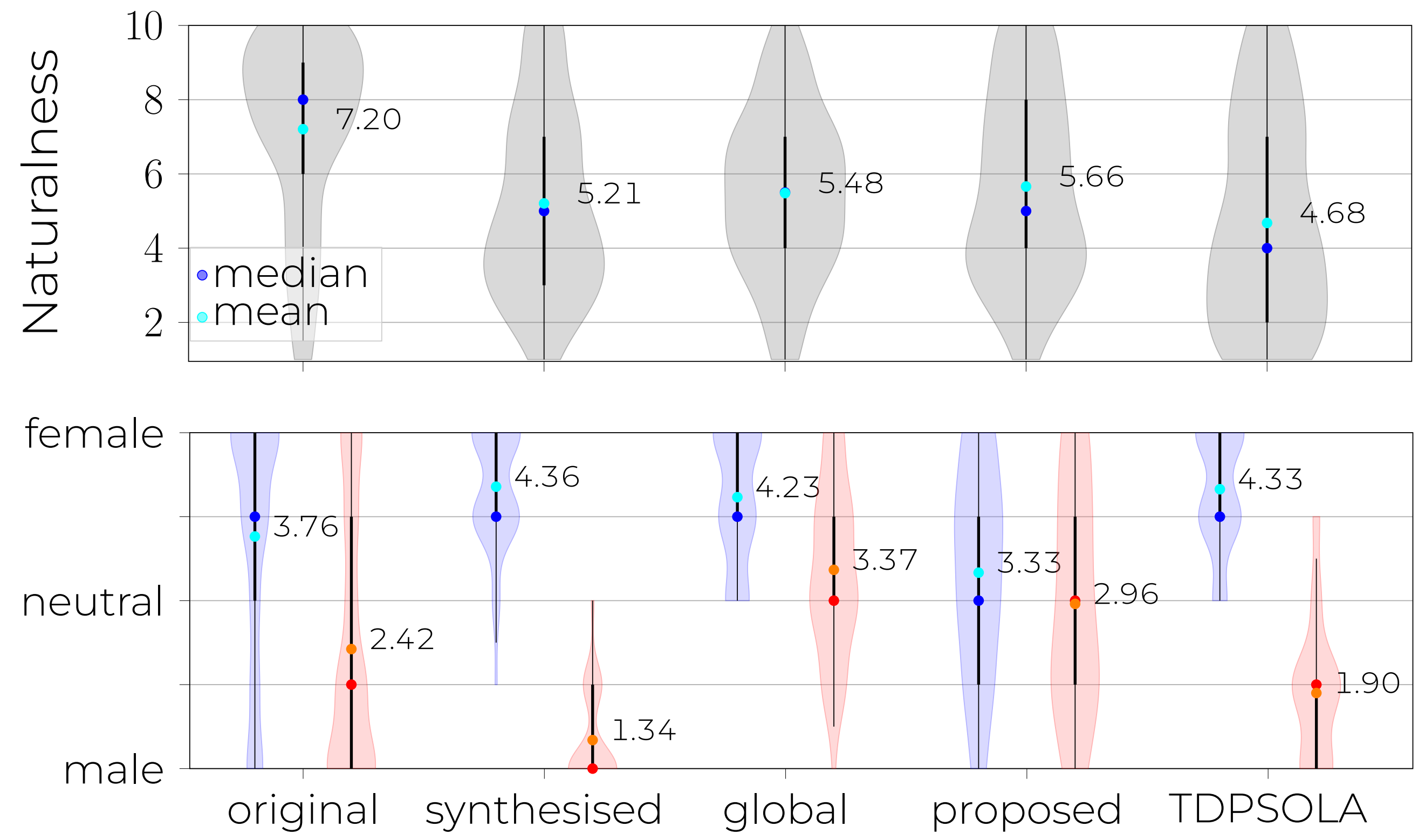


Figure: Listening test results. Violin plots of perceived speech naturalness (top). Violin plots of perceived speaker’s sex (bottom), blue for female and red for male; blue and red dots show medians, cyan and orange dots show means.

Conclusion

We proposed to reduce the speaker’s sex-related information in speech utterances by applying an affine transformation on the f_0 and the zero-LLR-based protection on the speaker embeddings.

The protection ability has been confirmed with automatic sex recognition and listening tests.