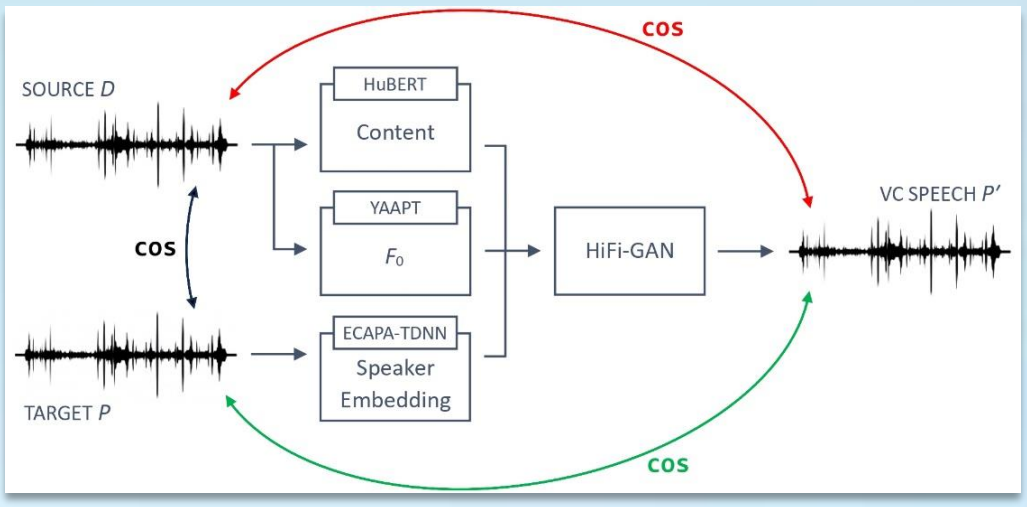
Quantifying Source Speaker Leakage in One-to-One Voice Conversion

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

To what extent can we **quantify confidence** about a **source speaker's ID** with **one-to-one voice conversion**? We compare **information leakage** for a range of **speaker characteristics**, in a 'worst-case' white-box scenario.

VC SYSTEM AND METHOD OF MEASUREMENT



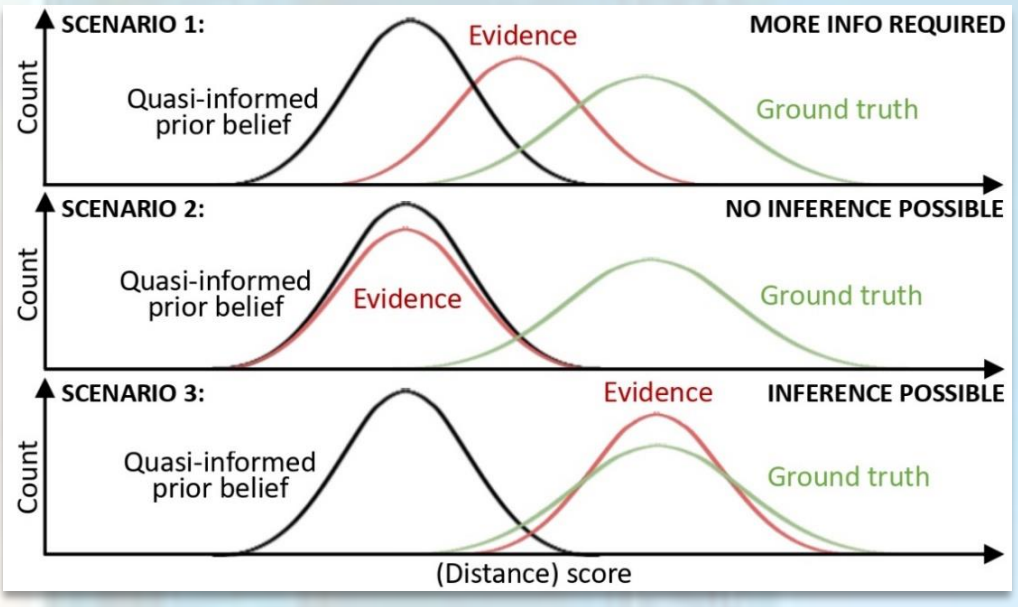
We use an SSL-based VC system with a neural speaker encoder to explore how **source speaker's information leakage** changes when VC is performed under both ideal and adverse conditions, with **mismatches** of **accents** and **recording environments**.

In an ideal VC system, the source speech provides only linguistic information. However, particularly with **Self-Supervised Learning (SSL)** architectures, a variety of other **non-linguistic information** may be used to inform a source speaker's identity.

Following extraction of **speaker embedding** vectors from the source, target and voice-converted speech, the distribution of their cosine distances can be evaluated using Earth Mover's Distance (EMD), an **interpretable metric** that expresses the minimum cost of transforming one distribution P into another Q: the higher the Wasserstein distance, the higher the EMD. Where $d_{i,j}$ is the ground distance between clusters p_i and q_j, and $f_{i,j}$ the flow between p_i and q_j that minimizes the overall cost:

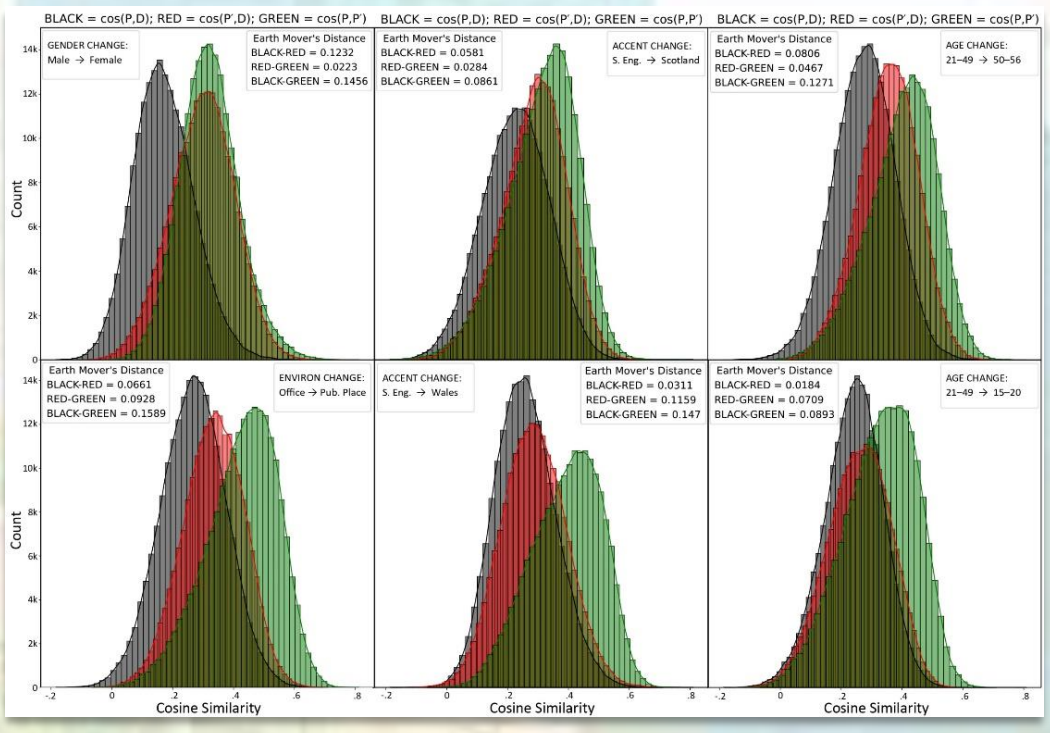
$$\mathrm{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$

INFORMATION-THEORETIC BASIS



We can now **quantify our confidence** to **infer source speaker characteristics** due to information leakage, via the distributional shift of evidence distribution ***R*** from ground truth ***G*** towards prior belief ***B***. In other words, we may define leakage L in terms of the proportional difference of EMD(***B***, ***R***) and EMD(***G***, ***R***) to EMD(***B***, ***G***):

EXAMPLES OF DISTRIBUTIONAL SHIFTS



$$L = \frac{\mathrm{EMD}(\mathbf{cos(source, target)}, \mathbf{cos(target, VC)})}{\mathrm{EMD}(\mathbf{cos(target, VC)}, \mathbf{cos(source, VC)})}$$

...where L tends towards 0 when no inference is possible, and increasing values give higher confidence to perform inference.

IN ANSWER TO THE OPENING QUESTION: A CASE STUDY

using speech corpora SPEECON (n = 194) and VCTK (n = 43):

[left] A **'sliding scale'** of how the choice of source speaker characteristics results in greater (or lesser) **interpretable data leakage**. More 'leaky' mismatched source speaker attributes (e.g. gender) result in greater interpretable leakage, while others (e.g. age < 21) result in lesser interpretable leakage.

Providers of synthetic voices must fulfil **legal and moral obligations to protect the identities** of their source speakers: methods to **dampen information leakage** or obfuscate such identifying features must be pursued.

Scott Wellington
University of Bath
Bath, United Kingdom
sdlw20@bath.ac.uk

Xuechen Liu
National Institute of Informatics
Tokyo, Japan
xuecliu@nii.ac.jp

Junichi Yamagishi
National Institute of Informatics
Tokyo, Japan
jyamagis@nii.ac.jp

UNIVERSITY OF BATH
UK Research and Innovation
art-ai
SpeakUnique