Revisiting Speech Content Privacy

Jennifer Williams[1], **Junichi Yamagishi**[2], **Paul-Gauthier Noé** [3], **Cassia Valentini-Botinhao**[1], **Jean-François Bonastre**[3]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]National Institute of Informatics, Japan
[3]University of Avignon, France
j.williams@ed.ac.uk

Abstract

In this paper, we discuss an important aspect of speech privacy: *protecting spoken content*. New capabilities from the field of machine learning provide a unique and timely opportunity to revisit speech content protection. There are many different applications of content privacy, even though this area has been under-explored in speech technology research. This paper presents several scenarios that indicate a need for speech content privacy even as the specific techniques to achieve content privacy may necessarily vary. Our discussion includes several different types of content privacy including recoverable and non-recoverable content. Finally, we introduce evaluation strategies as well as describe some of the difficulties that may be encountered.

Motivation

- Private content includes keywords or keyphrases such as named entities (places, dates, locations, organizations etc.), or financial and medical details
- Speech content privacy has been rooted signal-emitting devices that operate in physical spaces and conceal entire conversations.
- New use-cases for content privacy are emerging
- Explore new forms of content privacy as well as techniques and evaluation

Use Case #1 - Voice Storage Privacy

- Voice-enabled devices capture and store speech data, and in some cases also transmit it from the device to a larger database on a remote computing server

Use Case #2 - Speech Compression and Transmission

- Speech must be compressed for use in mobile phones, internet voice calling, and television broadcast, among others
- Law court testimony, emergency calls, and policing: protect the witness identity and hide or mask different levels of information, from voice identity to linguistic content

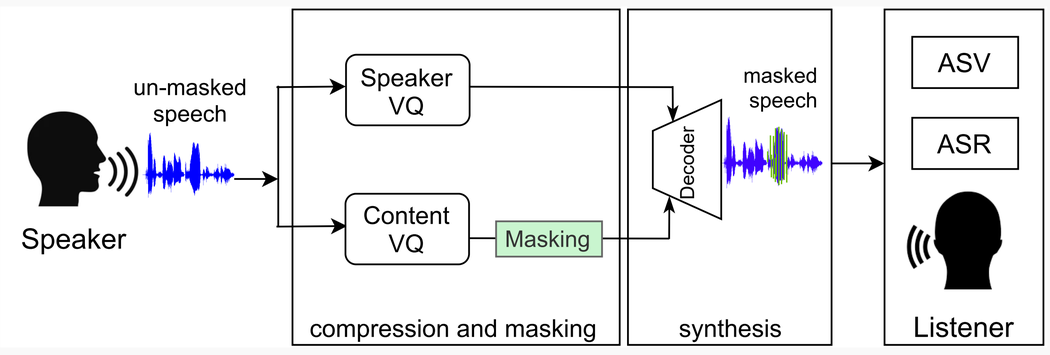
Use Case #3 - Speech and Speaker Recognition

- ASR requires that unmasked words remain highly intelligible
- ASV is a common application where the speaker information needs to be preserved when using content masking at the same time
- Content masking can be used for speech and speaker recognition if it does not alter speaker information or cause confusion between speakers

Use Case #4 - Voice-Enabled Assistive Technology

- Responses from Siri or Google may contain sensitive information that users do not want people nearby to overhear
- Blind users of screen reading technology do not enjoy the same expectations of privacy as sighted people when browsing the internet
- Develop a special earbud customized for blind users, while allowing for other important sounds in the environment

Content Privacy Approaches

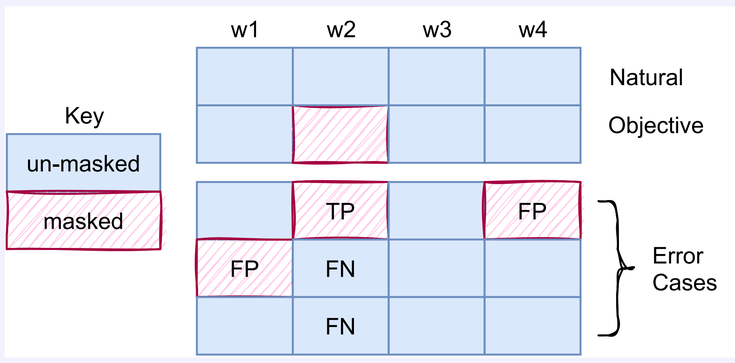


- Speech signal disentanglement stands out as a promising overall approach
- Main benefits of disentanglement: allows content to be modified separately from other informational factors
- Perform content privacy while signla is compressed
- Compression is beneficial for privacy related to transmission or broadcast

Evaluation: Task-Based

- ASV: ensure that speaker information is not altered
  Measure: EER, FRR, FAR
- ASR: ensure high intelligibility for un-masked surrounding words
  Measure: WER, PER

Evaluation: High-Level



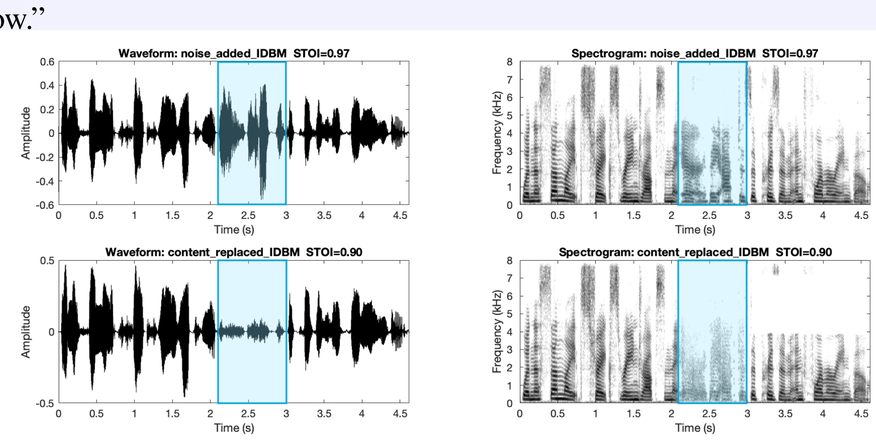
$$MER_{utt} = \frac{\alpha FN + \beta FP}{TN + TP + FN + FP} \quad (1)$$

MER requires high-quality time alignments in order to perform the calculations. For reversible masking, MER can be adapted to measure how much content is recoverable when the mask is reversed

- True negative (TN) word is correctly unmasked
- True positive (TP) word is correctly masked
- False negative (FN) word is incorrectly unmasked
- False positive (FP) word is incorrectly masked

Evaluation: Low-Level

"When sunlight strikes raindrops in the **[air they act as a]** prism and form a rainbow."



Waveform and spectrogram to compare ideal binary mask from two different content-masking approaches: additive noise (top) and replacement with noise (bottom). In each case, target content to be masked is highlighted by the box.

- Top: created by *adding* a temporally-modulated speech-shaped noise masker to the signal, to mask a target phrase, and then an ideal binary mask was calculated and applied to attempt to recover the speech
- Bottom: created by *replacing* a target phrase with the same type of noise.

Discussion and Future Work

- Speech disentanglement is a promising overarching approach for content privacy
- Content privacy will have a large impact on society since privacy concerns influence how people adopt new voice technologies
- Privacy controllability would allow users to have different privacy features in different settings, such as at home versus in public, or depending on who is nearby, or the type of voice device being used

This work was partially supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and University of Edinburgh; and by a JST CREST Grant (JPMJCR18A6, VoicePersonae project), Japan.