Speech as Personally Identifiable Information

Isabel Trancoso

inesc id lisboa

TÉCNICO LISBOA

Outline

- State of the art in speech tech
  - Predictions from turn of the century surveys
  - Smart speech tech 20 years after
- Security & privacy in speech communication
  - Threats
  - Counter-threats
- INESC-ID's work
  - Privacy preserving speaker verification
  - Privacy preserving extraction of health-related paralinguistic info
- Raising awareness about security & privacy in speech tech
  - Survey questions 20 years after
  - Towards usable privacy

Progress & Prospects for Speech Technology: Results from Three Sexennial Surveys , Roger K. Moore, 2011

- ASRU surveys: 1997, 2003, 2009
- *Insert the year in which you estimate the statement will become true (use "X" to indicate "never")*'.

1997 Survey by Roger K. Moore

- *More than 50% of new PCs have dictation on them, either at purchase or shortly after.*
- *Most telephone Interactive Voice Response systems accept speech input (and more than just digits).*
- *TV closed captioning is automatic and pervasive.*
- *Voice recognition is commonly available at home (e.g. interactive TV, control of home appliances and home management systems).*
- *Automatic airline reservation by voice over the telephone is the norm.*
- *Speech recognition accuracy equals that of the average (individual) human transcriber.*
  - *Predictions: 2020 (1997), 2030 (2003), 2035 (2009)*

1997 Survey by Roger K. Moore

- *It is possible to hold a telephone conversation with an automatic chat-line system for more than 10 minutes without realising it isn' t human*.
- *Voice-enabled command, control and communication in cars becomes as common as intermittent wiper, power window or power door lock*.
- *No more need for speech research*.
- *A leading cause of time away from work is being hoarse from talking all the time, and people buy keyboards as an alternative to speaking*.
- *Public proceedings (e .g. courts, public inquiries, parliament etc.) are transcribed automatically.*
- *First legal case in which a recording of a person's voice is thrown out because it cannot be proved whether a computer or a person said it.*
  - Predictions: 2020 (1997), 2020 (2003), 2030 (2009)

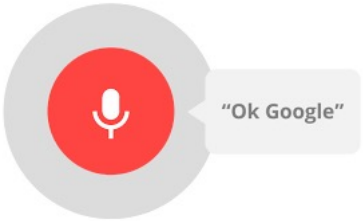
2003 Survey by Roger K. Moore

- *The majority of text is created using continuous speech recognition*.
- *Telephones are answered by an intelligent answering machine that converses with the calling party to determine the nature and priority of the call*.
- *Most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face)*.
- *Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language*.
- *Most interaction with computing is through gestures and two-way natural-language spoken communication*.
- *Pocket-sized listening machines are commonly available for the hearing impaired*.
- *The majority of automatic speech recognition systems have completely abandoned the HMM paradigm for acoustic modelling*.
  - Predictions: 2040 (2003), 2033 (2009)
- *The majority of automatic speech recognition systems have completely abandoned the n-grams paradigm for language modelling*.
  - Predictions: 2100 (2003), 2045 (2009)

2009 Survey by Roger K. Moore

- *Most information access and search using mobile phones are done through speech recognition and synthesis (e.g., web search, SMS).*
- *Mobile phones are used to control and monitor home appliances remotely using speech (e.g., remote access to DVR, recording programs, TV).*
- *Most multilingual people communicate with each other through speech to speech translation at any time using their mobile device.*
- *Number of speech-enabled applications created within the mobile ecosystem (e.g., Apple store, RIM, Android, etc) reaches 1 million.*
- *All mobile devices have built-in speech recognition capability.*
- *Mobile speech applications generate a $10 billion in revenue.*
  - *Predictions: 2020*

Predictions about smart speakers

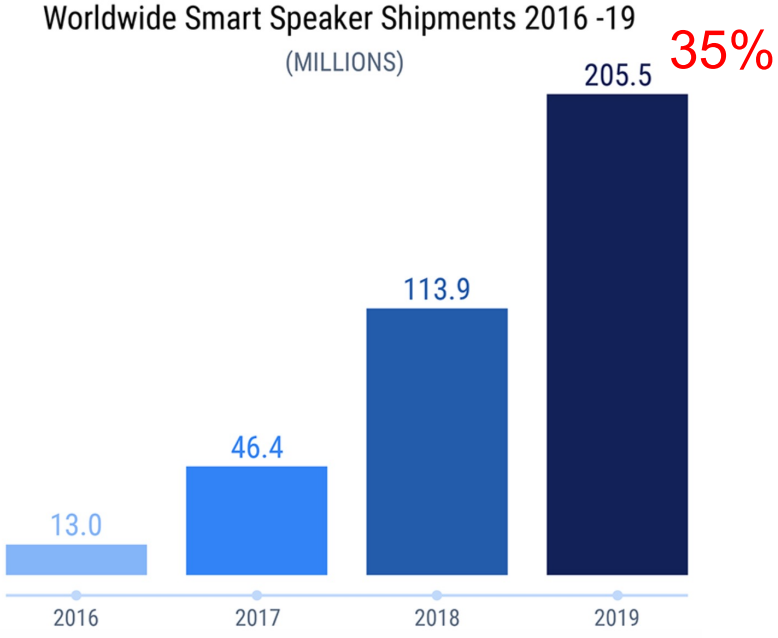


amazon alexa

Baidu Research

DRAGON NATURALLYSPEAKING

NUANCE



With the advent of deep learning …

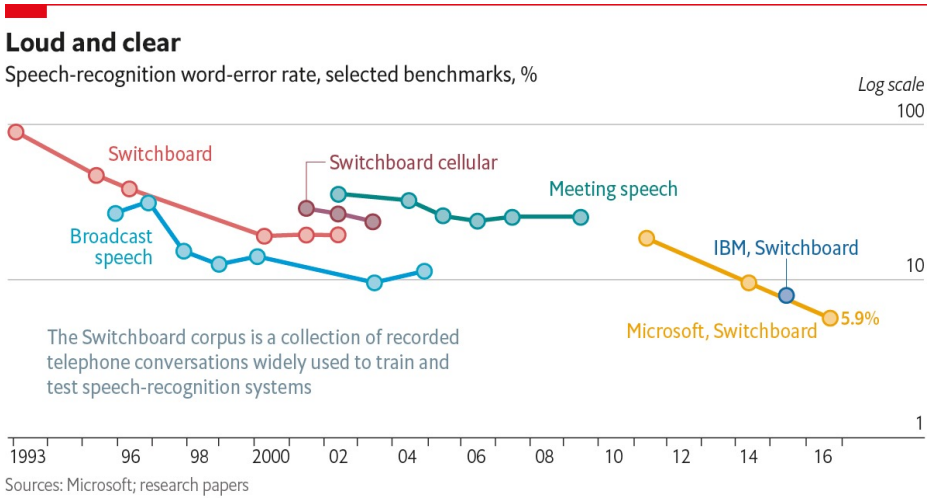
TÉCNICO LISBOA

SOTA: Speaker Representations

- Speaker embeddings (x-vectors)
- Datasets of over 7,000 speakers
- Automatic Speaker Verification with EER ~3%

SOTA: Automatic Speech Recognition

- From GMM-HMM to DNN-HMM and fully E2E ASR
- Audio augmentation, transfer learning, multi-task learning …
- WER ~ 4% for read speech, 3x for conversational speech
- Typically run **in the cloud**.



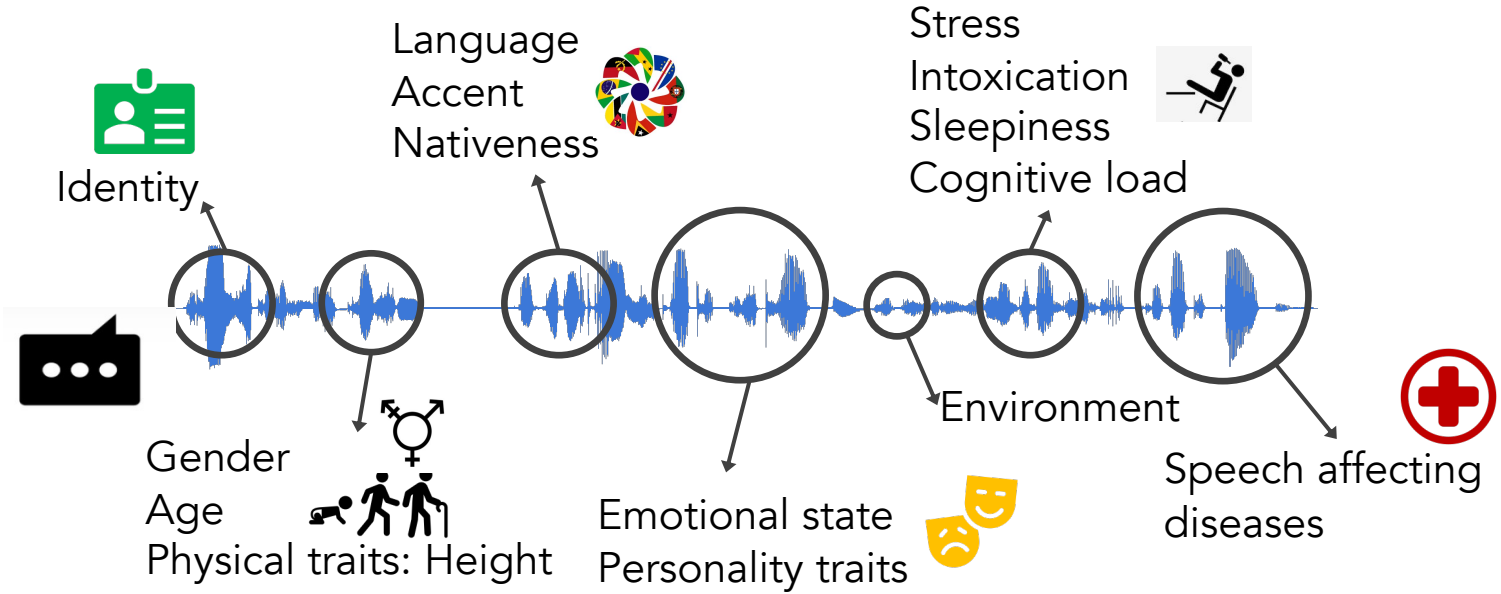
SOTA: Text-to-Speech Synthesis

- Traditional Vocoder → Neural Vocoder, taking as input time-frequency spectrogram representations
- Encoder-decoder architectures, with attention mechanisms mapping the linguistic time scale to the acoustic time scale
- Multi-speaker TTS systems, leveraging speaker embeddings.
- Zero-shot TTS, with only a few seconds (SC-Glow-TTS)
- MOS ~4
- Typically run **in the cloud**.

SOTA: Voice Conversion

- Traditional VC approaches:
  - analysis and feature extraction, mapping, and reconstruction
- Neural vocoders trained jointly with the mapping module and even with the analysis module to become an end-to-end solution
- Disentanglement of speaker and linguistic contents
  - variational auto-encoder schemes
  - content encoder learns a latent code from the source speaker
  - speaker encoder learns the speaker embedding from the target speaker speech.

SOTA – Speaker Profiling



Speech affecting diseases

- Much larger range than so-called speech & language disorders, e.g.:
  - Sigmatism, Stuttering
- Diseases that concern respiratory organs, e.g.:
  - Obstructive Sleep Apnea (OSA), Common Cold, COVID-19
- Mood disorders, e.g.:
  - Depression, Anxiety, Posttraumatic stress disorder (PTSD), Bipolar Disease
- Neurodegenerative diseases, e.g.:
  - Parkinson's disease (PD), Alzheimer's disease (AD), Huntington's disease
  - Amyotrophic lateral sclerosis (ALS)

- Datasets
  - Collection in clinical facilities, lack of longitudinal studies, different conditions
  - Crowdsourced collection (e.g. COVID-19, CLAC)
  - In-the-wild collection (e.g. WSM)

Vlogs

Speech affecting disease	UAR
Depression	0.810
PD	0.756
OSA	0.812

Security and Privacy in Speech Communication

- Almost absent from previous predictions
- Remote servers
- ISCA Special Interest Group SPSC
  - Interdisciplinary research: speech processing, user-interface design, law, cryptography, cognitive sciences, medical sciences.
  - An increasing number of devices are connected to the Internet and feature a microphone. These devices can provide many useful services such as personal speech assistance and security monitoring (e.g. banking, call centres and medical services). This trend however exposes users to an increasing number of threats.
  - The risks of breaches increase with the number of locations where data is processed and stored, the amount of data, the processing methods, storage formats and with whom the data is shared.

Voice Impersonation and Spoofing Attacks

- Voice impersonation attacks:
  - Human-based voice impersonation
  - Replay-based attacks
  - **Speech synthesis / voice conversion attacks**

- Spoofing: undermine voice-based verification systems
  - ASVspoof – Automatic Speaker Verification and Spoofing Countermeasures Challenge (2015, 2017, 2019, 2021)
  - ADD 2022 - Audio Deep synthesis Detection challenge (Mandarin)
    - Fully fake utterances with various real-world noises and background music effects
    - Partially fake utterances
    - Generate and detect attack utterances

Deep Fakes

- Past technology enabled simple manipulation of audio/video recordings
  - Cropping
  - Modifying speech rate
  - Modifying pitch
  - Could be used to manipulate public conscience

Deep Fakes

Collaboration with M. Figueiredo & A. Oliveira

Approaches to Privacy Preservation for Speech

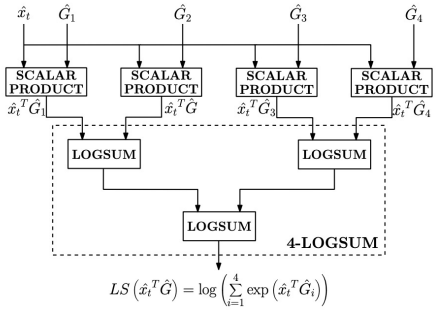
- Anonymization
  - Suppressing PII in the speech signal, leaving all other attributes intact, i.e. concealing the speaker's voice identity while protecting linguistic content, paralinguistic attributes, intelligibility and naturalness.
  - Voice Privacy Challenge 2020
    - Baseline system: x-vectors, ASR acoustic models, F0, neural source-filter model
    - Other approaches: noise addition, speech transformation, voice conversion / speech synthesis, adversarial learning, disentangled representation learning
  - Voice Privacy Challenge 2022
    - Additional baseline systems: Unified HiFi-GAN neural source-filter model and signal manipulation using pole position shifting
    - Stronger attacker models (ASV trained on anonymized data)
    - Complementary metrics: objective (WER, EER, pitch correlation, gain of voice distinctiveness) and subjective (intelligibility, naturalness, speaker verifiability)

Approaches to Privacy Preservation for Speech

- Encryption
  - Homomorphic Encryption (HE)
  - Secure Multi-Party Computation (MPC)
  - Secure Modular Hashing (SMH)
- Tasks
  - Privacy preserving speaker verification
  - Privacy preserving extraction of health-related paralinguistic info

Privacy Preserving Speaker Verification

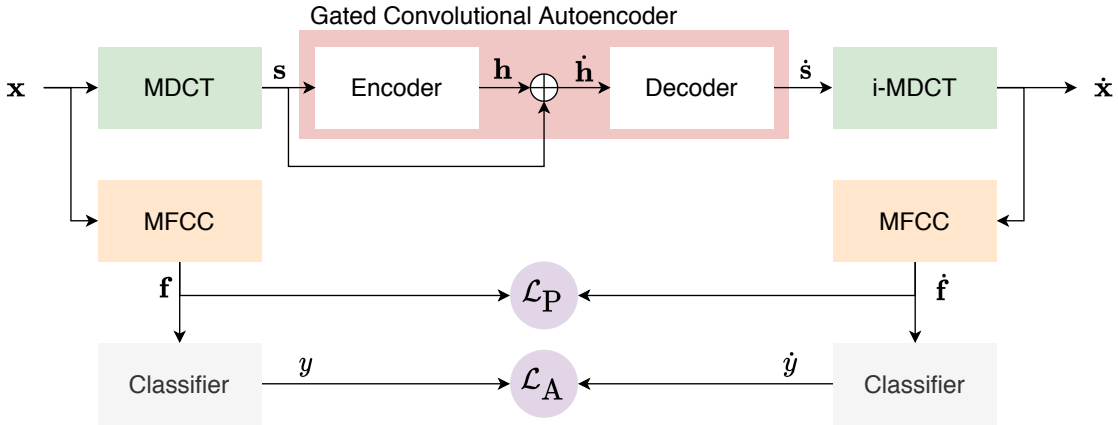
- Baseline ASV: UBM-GMM
- Remote privacy-preserving speaker verification system:
  - Neither the system observes voice samples / speech models from the user
  - Nor the user observes the UBM of the server
- Approach: Garbled Circuits
  - Secure Function Evaluation framework
  - Two parties can compute a function on their combined inputs without revealing their individual inputs to one another.



Privacy Preserving Speaker Verification

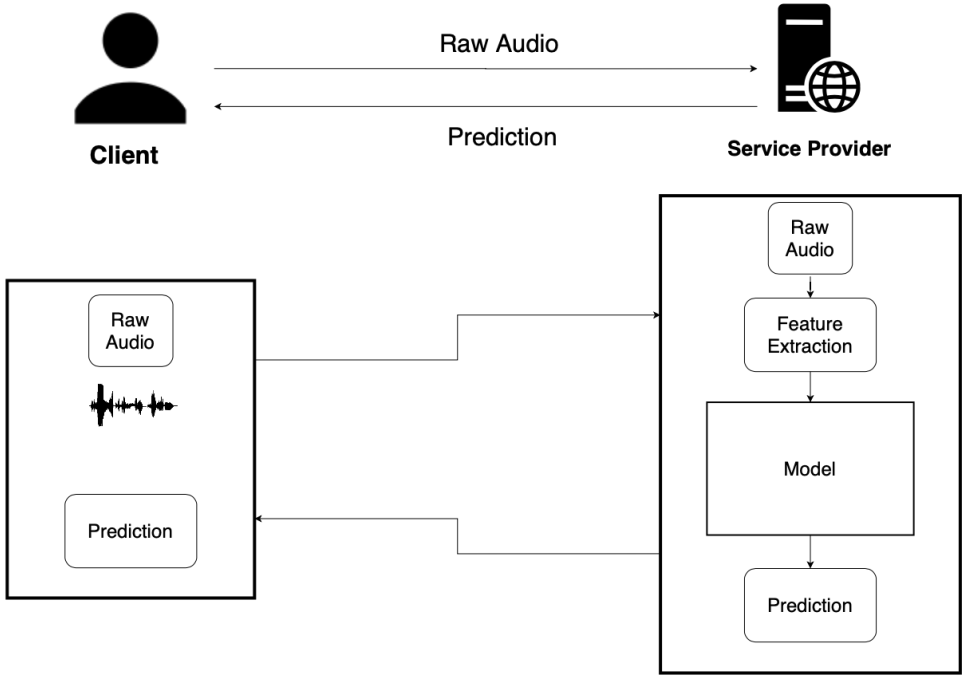
- Recent approaches:
  - Nautsch et al. 2018, Encryption of the user's i-vector using the public key of the authentication server
  - Mtibaa et al. 2021, Binarization of i-vectors and x-vectors
- Computation of the speaker embeddings by the user
  - Knowledge of the extraction model may help attackers in crafting adversarial examples able to mislead the authentication system.
  - Implies disclosing by the service provider one of the most valuable components in the system.
- Alternative: extract speaker embeddings while keeping both the speaker's voice and the service provider's model private
  - Secure Multi-Party Computation
  - Reasonable trade-offs between security and computational cost.

Speaker Verification Attack

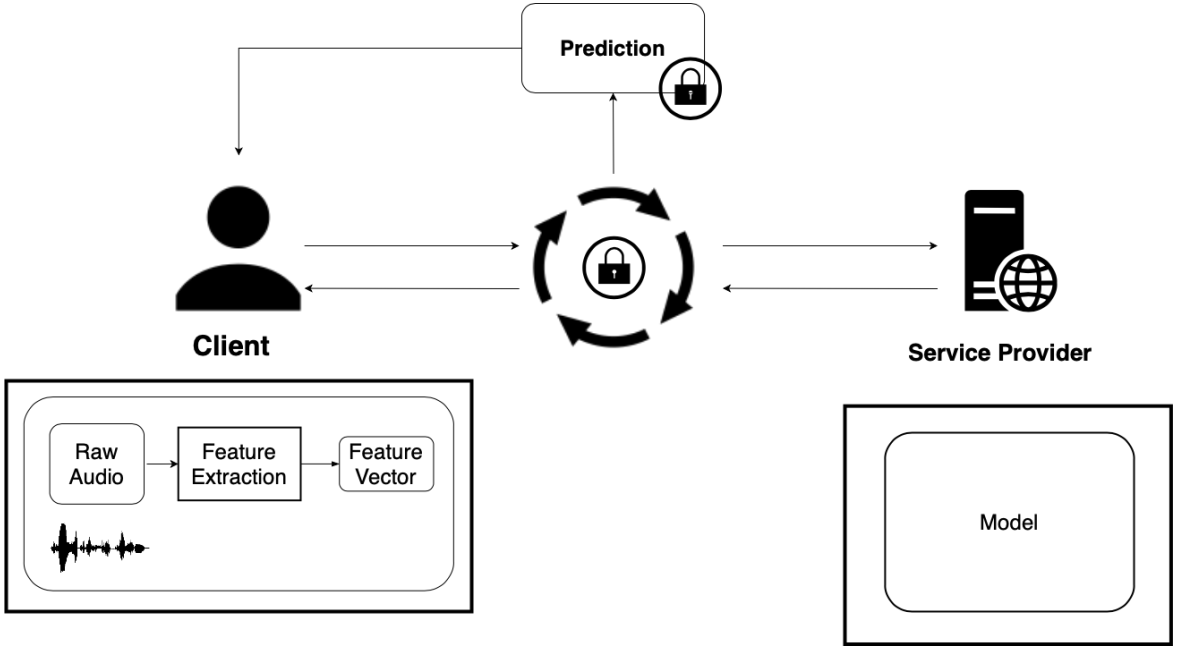


- Generate highly imperceptible adversarial disturbances against a speaker id model
- Multi-objective loss function:
  - Hinders speaker identification performance
  - Accounts for human perception (frame-wise cosine similarity of MFCCs)
- Experiments with a 250-speaker identification x-vector network
- Highly imperceptible adversarial perturbations (PESQ scores above 4.30)
- Success rate of 99.6% and 99.2% in misleading the speaker id model (for untargeted and targeted settings, respectively)

Machine Learning as a Service for Paralinguistic Tasks



Privacy-preserving MLaaS for Paralinguistic Tasks

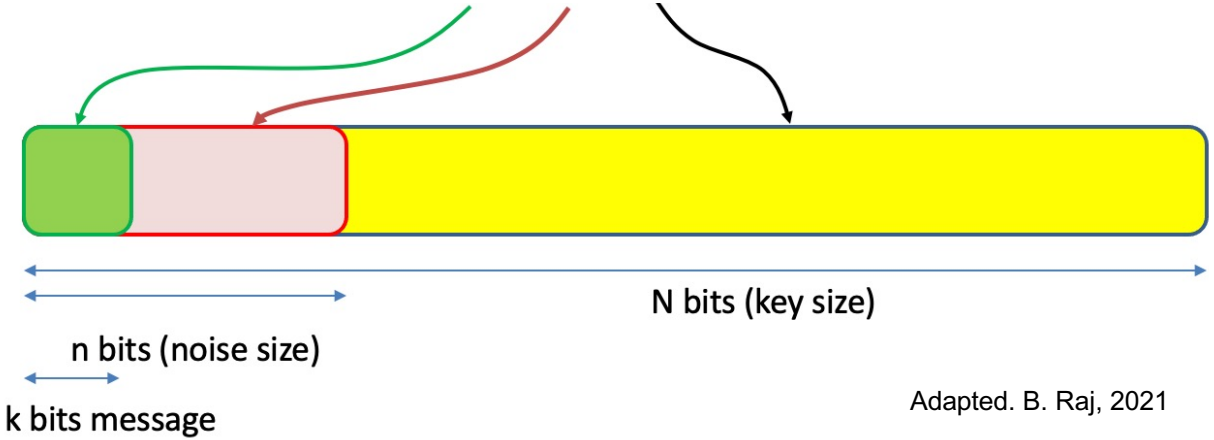


Privacy-preserving Neural Networks (eNNs)

Homomorphic Encryption (HE):

$$E(x) \otimes E(y) = E(x \times y)$$
$$E(x) \oplus E(y) = E(x + y)$$

$$c = \boldsymbol{b} + 2p + kx$$



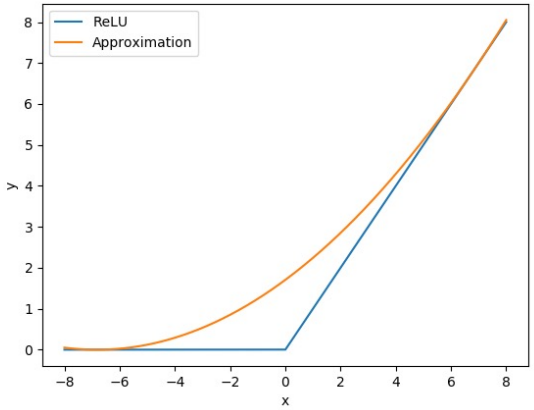
Adapted. B. Raj, 2021

$$\boldsymbol{b} = (c \bmod k) \bmod 2$$

- All operations in the NN are replaced by their HE counterparts (Cryptonets, CryptoDL).
- HE-based solutions rely on noise to hide the plaintext.
- "Noise" grows with homomorphic operations → Heavy computational load.

Privacy-preserving Neural Networks (eNNs)

- Non-linear activation layers are replaced by polynomial approximations.
  - May lead to less accurate models.
- Batch Normalization layer is introduced before each activation, to ensure inputs fall within the convergence interval.
  - Requires discretized weights & inputs;
  - May also lead to accuracy degradation.



Privacy-preserving Neural Networks (eNNs)

Accuracy

- For Cold, Depression, and Parkinson's Disease, experiments using eGeMAPS and encrypted NNs showed small degradation.
  - At a speaker level, degradation becomes negligible

Computational costs

- Predictions take much longer than in an unencrypted context:
  - Seconds vs microseconds (for isolated predictions)
  - Milliseconds (using batching for simultaneous encryptions)

But

- Network architecture is limited by *noise* growth, scaling and computational complexity

Privacy-preserving Support Vector Machines (eSVMs)

Radial Basis Function (RBF) kernel

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}_i') = \exp(-\gamma d_E^2(\boldsymbol{x}, \boldsymbol{x}_i'))$$

$$\mathrm{k}(d_H') = \exp(-\frac{\gamma}{M^2} {d'}_H^2)$$

Secure Modular Hashing: Distance preserving hash function

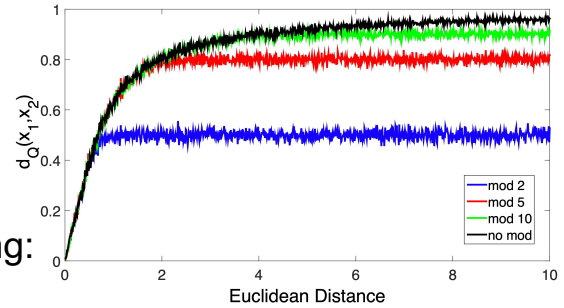
$$d_H(x, y) = \frac{1}{M} \sum_{i=0}^{M} x_i \oplus y_i$$

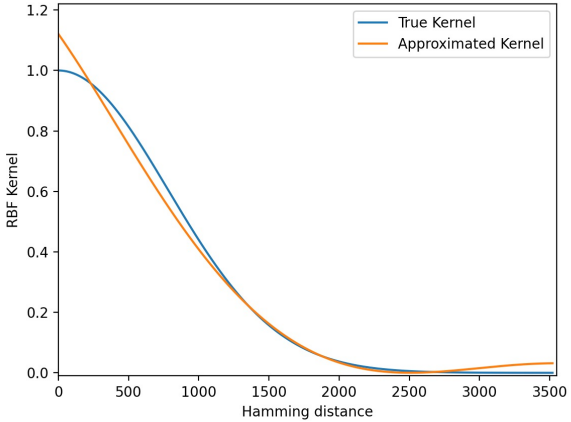
$$\hat{y}(\boldsymbol{x}) = \mathrm{sign}(w_0 + \sum_{i=0}^{n} \alpha_i y_i k_H(\boldsymbol{x}, \boldsymbol{x}_i'))$$

Garbled Circuits

Homomorphic Encryption

Secret Sharing + Homomorphic Encryption





Privacy-preserving Support Vector Machines (eSVMs)

Accuracy

- For Obstructive Sleep Apnea and Parkinson's Disease, experiments using knowledge-based features and encrypted SVMs showed no relevant accuracy degradation relative to non-encrypted.

Security and Computational Cost

- The security of this method comes from the security of the underlying HE and MPC protocols.
- Single prediction: ~600ms and 3MB of bandwidth

Privacy-preserving paralinguistics

- Proof-of-concept of how paralinguistic health-related tasks can be made secure through the combination of HE and MPC methods.
- eNN
  - Discretizing an eNN can be done with minimal accuracy degradation.
  - Not scalable to deeper networks, unless we alternate between protocols.
- eSVM
  - SMH may be used to replace ED with the much lighter computation of HD, reducing the computational cost of the eSVM with the RBF kernel.

But

- The computational and bandwidth cost of these methods is still much higher than their in-the-clear counterparts!
- The computational toll due to feature extraction is on the client's side.

Outline

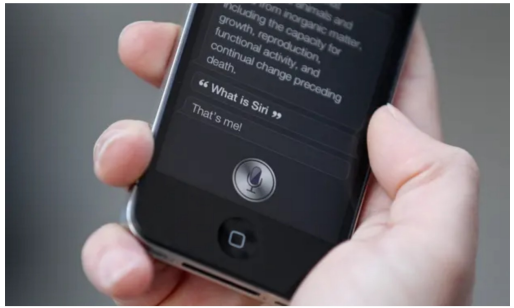
- State of the art in speech tech
  - Predictions from turn of the century surveys
  - Smart speech tech 20 years after
- Security & privacy in speech communication
  - Threats
  - Counter-threats
- INESC-ID's work
  - Privacy preserving speaker verification
  - Privacy preserving extraction of health-related paralinguistic info
- **Raising awareness about security & privacy in speech tech**
  - Survey questions 20 years after
  - Towards usable privacy

Speech as PII (Personally Identifiable Information)

Amazon’s Alexa recorded private conversation and sent it to random contact

The "Big Brother 2019 Award" for privacy and data collection violation this year goes to Amazon (among others)

Speech assistant: Google stops analysis - for now



Apple halts practice of contractors listening in to users on Siri

Siri and Alexa could become witnesses against you in court some day
https://sdgln.com/news/2016/12/29/siri-and-alexa-could-become-witnesses-against-you-court-some-day

LaLiga fined for soccer app's privacy-violating spy mode
https://techcrunch.com/2019/06/12/laliga-fined-280k-for-soccer-apps-privacy-violating-spy-mode/

2021 Survey

Speech Synthesis/Voice Conversion

- *Are you aware that a handful of utterances is enough to build your own synthetic voice?*
- *Is it enough to tag a synthetic voice as “deep fake” to make its dissemination legal?*
- *Who is responsible when a digital voice is made to say things that the original speaker would find offensive or untrue?*
- *If we bring to life a voice of the past, who owns the words?*

Speech-to-Speech-Machine Translation with Voice Conversion

- *Would you use an S2SMT system that translates your utterances in your native language to a target language using your own voice (knowing that the product vendor stores your speaker representation) ?*

Computer Assisted Language Learning with Voice Conversion

- *Would you use an automatic pronunciation training tutor that provides you with reference utterances in a foreign language in your own voice (knowing that the product vendor stores your speaker representation)?*

2021 Survey

Mining paralinguistic and extra-linguistic information from speech

- *Do you feel that humans can be profiled from their voice?*

Personal Voice Assistant

- *Do you trust your PVA?*
- *Do you feel that your PVA is spying on you?*
- *Do you know how to turn off the always listening mode?*

Speech product vendor subcontracted by a bank / clinic / enery provider / telecom provider, insurance company, restaurant, call center, etc.

- *Are you aware that your audio data may be stored in servers of subcontracted speech product vendors ? And for how long?*

Audiovisual recordings of interviews for research studies

- *Is occlusion of facial features enough for anonymization?*
- *Should voices be anonymized?*

Towards usable privacy

- Customization to the user's needs
- Utility versus privacy trade-off
- Disruption of the business models of many current service providers
- Reduce computational costs and bandwidth in transmission

Thank you!
Obrigada!
Gracias!

Organizing Committee

Bhiksha Raj, Francisco Teixeira,
Alberto Abad, Catarina Botelho, Joana Correia
Roger K. Moore, Andreas Nautsch, Tom Backstrom