# Achieving Transparently Synthetic Speech with a Speaker Cocktail

*Jessica Huynh*, Jiatong Shi*, Xuankai Chang*, Jeremiah Milbauer*, Shinji Watanabe*

Carnegie Mellon University

{jhuynh, jiatongs, xuankaic, jmilbaue, swatanab}@cs.cmu.edu

## Abstract

The rapid progress of speech synthesis technology indicates that synthesis will soon achieve human parity in fluency and naturalness, and introduce new opportunities for misuse. We demonstrate how generated high-quality human-like speech can also be used to address these risks. This paper introduces a voice conversion technique that incorporates not only anonymization, but also disclosure-of-synthesis into the same speech signal through a novel decoding approach called "Cocktail Speaker Decoding" (CSD). Without re-training the network, CSD dynamically changes speaker identities by altering each synthesized utterance to contain a shift in speaker identity throughout the clip. Our experiments on LIBRITTS, using both textual and HUBERT discrete-unit representations, show that the proposed framework can synthesize high-quality, anonymous, and transparently synthetic speech – as judged by word-error-rate, equal-error-rate, and subjective human evaluation.

**Index Terms**: transparently synthetic speech, speaker cocktail, speech anonymization, speech privacy, speech synthesis

## 1. Introduction

Speech synthesis is rapidly becoming more human-like, and coupled with the widespread proliferation of machine learning APIs, this development will create a number of new risks regarding the abuse of speech synthesis technologies. Attacks are already emerging. A corporate executive's synthesized voice was used to scam a subsidiary out of company funds [1]. A woman fabricated evidence in a court case with synthesized speech she had learned how to create online [2]. On a larger scale, policy analysts are already considering how voices shared publicly online could be collected by malicious actors to manipulate targets through impersonation [3].

Human perception of synthesized speech is of concern, as synthesized speech can be used to deceive real-world systems, and in controlled settings does not trigger suspicion or hesitancy without explicit reason [4]. In settings where a speech signal may already suffer from degradation, such as over the internet or phone, humans rate high-quality synthesized speech as equal or even more human-like than genuine human speech [5]. There are additional ethical questions about the treatment of human users, who may interact differently with human agents compared to artificial intelligence (AI) agents [6].

With legislative developments now requiring AI disclosure [7], it will also be important to develop systems that not only *allow* for disclosure, but also *guarantee* it when those systems are made public, while still maintaining naturalness and clarity. Merely labeling audio is not sufficient – the label could be easily removed, or forgotten. While previous work has explored watermarking a generated speech signal [8, 9], watermarking not discernible to the human ear merely achieves disclosure, but not the "clear, conspicuous" transparency required by law.

When building defenses that protect voices (anonymization), we must also prevent their deceptive use for human impersonation by developing approaches to transparent, i.e. clearly noticeable, disclosure. These should be general enough to evolve *alongside* state of the art speech generation technologies – rather than waiting for, and reacting to, the inevitable abuse.

A popular approach to anonymization, the use of single-speaker voice conversion, improves anonymization but not necessarily disclosure. Another issue is that audio clips may contain identifiers [10], and it is difficult to extract speaker agnostic representations without introducing noise [11]. To increase single-speaker anonymization, recent contributions modify the x-vector of the speaker [12, 13], using methods such as choosing a nearby speaker [14, 15], and calculating a mixture of x-vectors [16]. However, none of these techniques combine disclosure with speech anonymization.

Many methods are emerging for disclosure. One approach is to provide a warning label which discloses that a user is interacting with an AI agent. If systems provide explicit disclosure of this kind, consequences can be positive (humans can discuss more negative topics to a bot than a human [17] or feel less anxious [18]) and negative (humans may develop a reliance [19], and bots can build and exploit trust [20]). If systems do not, human users may engage in an ad-hoc Turing Test [21] to determine if the agent is human or machine – and in many of these cases, systems do not reliably state that they are not human [22].

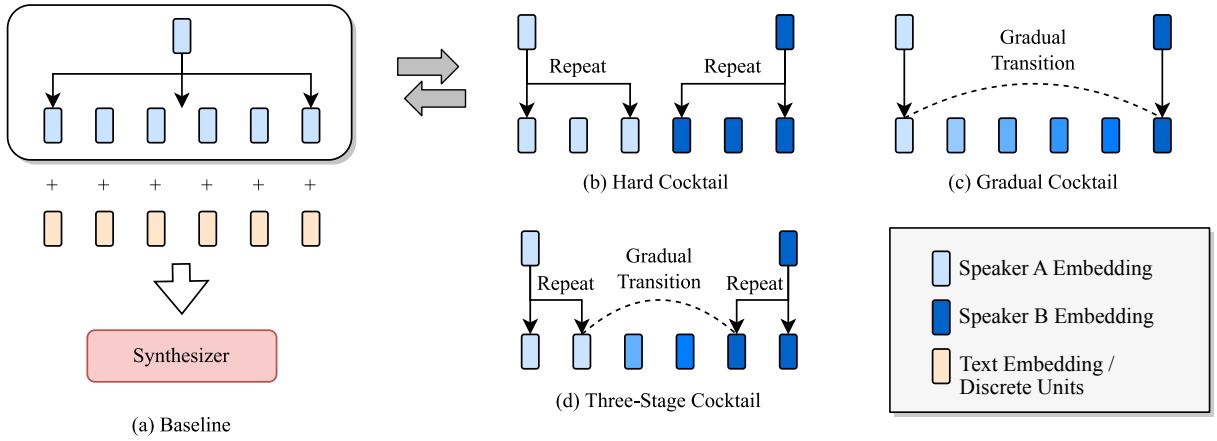
Thus, we propose a novel framework for *transparently* synthesized speech, called **cocktail speaker decoding** (CSD). On top of a voice conversion system, during decoding, the synthesized speaking style is continually modified by changing the speaker embedding given to the decoder, which results in a cocktail of ever-changing human voices within the same utterance. Without re-training the network, this results in speech that is natural, clean, and obviously synthesized. As the approach is agnostic to the voice conversion system used, it should remain a reasonable strategy to achieve both anonymity and transparency as speech synthesis progresses.

Our empirical studies on LIBRITTS [23] indicate that the CSD method is not only more *natural* according to subjective human evaluation, but also surprisingly outperforms single-speaker voice conversion both in terms of intelligibility and anonymization.

## 2. Voice Conversion for Speech Anonymization

Denote $\boldsymbol{X} \in \mathbb{R}^{1\times T}$ as a speech signal of length $T$ and $s_{\text{src}} \in \mathbb{R}^{d_s}$ as its $d_s$-dimensional speaker embedding. We define $\mathcal{F} : \mathbb{R}^{1\times T} \to \mathbb{R}^{1\times T'}$ as the function space that maps the speech signal $\boldsymbol{X}$ to another $\boldsymbol{X}'$ with length $T'$. Then, we denote the perceptible speaker-free information within $\boldsymbol{X}$ as

*Equal contribution




(a) Baseline (b) Hard Cocktail (c) Gradual Cocktail (d) Three-Stage Cocktail

Figure 1: *Cocktail speaker decoding: (a) Baseline with consistent speaker embeddings for decoding. (b) Hard cocktail: cocktail decoding with hard speaker embedding shift. (c) Gradual cocktail: cocktail decoding with gradual speaker embedding shift. (d) Three-stage cocktail: combination of hard cocktail and gradual cocktail decoding.*

$\boldsymbol{L} \in \mathbb{R}^{m \times n}$, where $m$ is the temporal length of $\boldsymbol{L}$ and $n$ is the feature dimension of $\boldsymbol{L}$ at each temporal frame. As noted in Eq. (1), the objective of speech anonymization is to find the best function $f^* \in \mathcal{F}$ that minimizes the probability of source speaker $s_{\text{src}}$ conditioned on the speech signal $\boldsymbol{X}$.

$$\text{argmin} \quad P(s_{\text{src}}|f(\boldsymbol{X})) \tag{1}$$

$$\text{subject to} \quad P(\boldsymbol{L}|\boldsymbol{X}) = P(\boldsymbol{L}|f(\boldsymbol{X})) \tag{2}$$

When applying voice conversion, the function $f(\cdot)$ can be decomposed into $r(g(\cdot))$, where $g : \mathbb{R}^{1\times T} \rightarrow \mathbb{R}^{m\times n}$ expects to remove the speaker information $s_{\text{src}}$ and $r : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{1\times T'}$ re-synthesizes the $s_{\text{src}}$-independent speech. The intermediates (i.e., the output of $g$) are expected to be speaker-free $\boldsymbol{L}$.

### 2.1. Cascaded Voice Conversion (CVC)

Cascaded voice conversion has shown to be a good candidate for privacy protection during which the speaker identity $s_{\text{src}}$ of a speech signal changes [24]. It applies an ASR system, $g$, for information filtering and a text-to-speech (TTS) model, $r$, for speech re-synthesis [25, 26]. Based on the formulation, the intermediate speaker-free information $\boldsymbol{L}$ is defined as one-hot textual representation $\boldsymbol{L}^w \in \mathbb{R}^{m^w \times n^w}$, where $n^w$ is the length of linguistic units (e.g., characters, phonemes, or words), and $m^w$ is the size of vocabulary. During the training, the TTS model follows the typical TTS training procedures. During inference, the TTS module conditions on a speaker representation $s_{\text{tgt}}$ that is different from source speaker's $s_{\text{src}}$.

### 2.2. Discrete Unit Voice Conversion (DUVC)

Recently, synthesizing speech from the discrete units obtained from HuBERT [27] has been shown to be feasible [28,29]. The state-of-the-art self-supervised learning representation model, HuBERT, is used to get the representations $\boldsymbol{H} \in \mathbb{R}^{m^h \times n^h}$ from the input speech $\boldsymbol{X}$. The discrete unit sequence $\boldsymbol{L}^h \in \mathbb{N}^{m^h \times 1}$ is generated by applying k-means clustering algorithm over the representations $\boldsymbol{H}$. Compared to CVC, all intermediate representations are now learned from data without supervision. To re-synthesize the speech, a modified HiFi-GAN [30] based vocoder is trained to generate the speech from the discrete units. The inference is similar to the CVC system; the vocoder conditions on a speaker representation $s_{\text{tgt}}$ and the discrete unit sequence to generate the target speaker's speech.

## 3. Cocktail Speaker Decoding

As discussed in Sec. 1, we argue the importance of adding obvious anonymization to synthesized speech while still keeping its naturalness. However, this is different from the training objectives in a voice conversion system as noted in Eq. (1). If the voice conversion system has a good performance over synthesized speech, it is likely to be difficult for listeners to notice that the speech is anonymized. On the other hand, if the system does not result in a reasonable quality, the speech will have the risk of losing speech intelligibility and naturalness.

To mitigate the aforementioned issue, we propose a decoding strategy, namely cocktail speaker decoding (CSD), during re-synthesis process $r$, shown in Fig. 1. In voice conversion systems introduced in Sec. 2.1, we adopt a synthesizer $r$ to re-synthesize the speech from its corresponding speaker independent information $\boldsymbol{L}$ and another speaker's embedding $s_{\text{tgt}}$. As illustrated in Fig. 1(a), the speaker embedding is repeated across time to align speaker independent information $\boldsymbol{L}$. However, in CSD, we introduce other unknown speakers and cocktail their embeddings with $s_{\text{tgt}}$. For simplicity, we only add one additional speaker information $s_{\text{tgt2}}$ compared to the baseline.

We propose three approaches for CSD in Fig. 1(b)-(d): hard cocktail, gradual cocktail, and three-stage cocktail.

**Hard Cocktail**: as defined in Sec. 2.1, the speaker independent information $\boldsymbol{L} \in \mathbb{R}^{m\times n}$ has $m$ frames. Instead of repeating $s_{\text{tgt}}$ over all $m$ frames, the hard cocktail applies $s_{\text{tgt}}$ on $\boldsymbol{L}_{1:\frac{m}{2}}$, which covers the first half frames of $\boldsymbol{L}$. Then it uses $s_{\text{tgt2}}$ on $\boldsymbol{L}_{\frac{m}{2}:m}$, which is the second half frames of $\boldsymbol{L}$. An illustration of hard cocktail is shown in Fig. 1(b).

**Gradual Cocktail**: The gradual cocktail is illustrated in Fig. 1(c). The method applies $s_{\text{tgt}}$ and $s_{\text{tgt2}}$ as the first and last speaker vectors injected to $\boldsymbol{L}$, respectively. While in the middle, we use a linear interpolation by defining the gradual change

factor as $\frac{s_{\text{tgt2}} - s_{\text{tgt}}}{m-1}$. Then for the $t^{\text{th}}$ frame $\boldsymbol{L}_t$ in $\boldsymbol{L}$, we have:

$$s_t = s_{\text{tgt1}} + \frac{t \cdot (s_{\text{tgt2}} - s_{\text{tgt}})}{m-1} \tag{3}$$

**Three-stage Cocktail**: in Fig. 1(d), for three-stage cocktail, we combine both the hard and gradual cocktail. $\boldsymbol{L}$ is divided into segments of equal length, $\frac{m}{3}$, where for the first and last segments, we use repeated $s_{\text{tgt}}$ and $s_{\text{tgt2}}$, respectively. For the middle section, we employ the gradual transition defined in Eq. (3).

# 4. Experiments

## 4.1. Datasets

We conduct our experiments with LIBRISPEECH [31] and LIBRITTS [23]. For the CVC system, we apply the whole training set of LIBRISPEECH to train the ASR model and train-clean-460 of LIBRITTS to train the TTS model for CVC and the vocoder for DUVC. We keep the original sampling rate of the dataset for model training for CVC (i.e., 16k Hz sampling rate for LIBRISPEECH and 24k Hz sampling rate for LIBRITTS), while the DUVC system uses 16k Hz sampling for both LIBRISPEECH and LIBRITTS. During objective evaluation, we downsample the CVC speech signals to 16k Hz to be comparable.

## 4.2. Experimental Setup

The CVC system adopts three models that are optimized separately. The input features of the system are log-Mel filter bank features that were extracted with an 8 ms frame shift and a 32 ms window length. For the ASR model, we use a pre-trained conformer-based encoder-decoder architecture trained with hybrid CTC/Attention from ESPnet [32–34]. For the TTS module, we use Fastspeech2 [35] with Conformer encoders. The text recognized from ASR is first converted into phonemes and then fed into the TTS module. We use x-vectors [36] as speaker information. During training, the x-vector of the source speaker (i.e., $s_{\text{src}}$ is concatenated to the encoder outputs to reflect speaker specifics. Detailed settings follow the configuration in ESPnet's LIBRITTS recipe [37,38]. A compatible HIFI-GAN vocoder from [39] is used to convert the Mel spectrogram from the TTS model into waveforms.

In DUVC, the pretrained HUBERT Base model[1] is used to generate the HUBERT representations from the 9th layer. A k-means clustering model is trained using the features from the train-clean-100 set with 100 clusters. Using the HUBERT and the k-means models, we can generate the discrete units for the whole training set. A HIFI-GAN vocoder taking the discrete units as input [29] is trained to generate the waveform speech.

For both CVC and DUVC, we consider five decoding strategies: (1) baseline: apply repeated speaker x-vector $s_{\text{tgt}}$, which differs from the source speaker, but is one of the speakers used for the cocktail methods, seen in Fig. 1(a). (2) x-vector perturbation: similar to [12], we perturb x-vector $s_{\text{tgt}}$ by adding noise sampled from a normal distribution. The strategy aims to project the x-vector from a real speaker to a fake speaker, so as to increase the awareness of "disclosure". (3-5) hard cocktail, gradual cocktail, and three-stage cocktail: the three cocktail strategies are illustrated in Fig. 1(b-d) and discussed in Sec. 3.

[1] https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt

## 4.3. Evaluation Metrics

We focus on three evaluation metrics: Objective word-error-rate (WER) for speech intelligibility, objective equal-error-rate (EER) for speaker verifiability, and subjective mean opinion score for speech naturalness. These are a subset of the metrics established by the Voice Privacy Challenge [40].

The main measure of speaker verifiability is equal-error-rate (EER). We calculate this by using a pre-trained Kaldi model for x-vectors trained on SITW [41]. We compare each of our configurations against the clean ground truth speech, and ground the calculation with target ground truth pairs, present in the test set of LIBRITTS, to determine if the method has successfully anonymized the ground truth utterance.

The generated speech, although not identifiable, should still be intelligible. Therefore, to measure how well the linguistic information is kept, we perform automatic speech recognition on the generated speech. We use the pre-trained ESPnet model on GIGASPEECH [42] to avoid training and testing on the same dataset [43]. The metric is word-error-rate (WER), which calculates the normalized distance between the predicted transcription and the ground truth text.

In addition to objective evaluations, we perform Mean Opinion Score (MOS) human evaluation [44] using Amazon Mechanical Turk (AMT). We choose a male and a female speaker from the LIBRITTS training set as $s_{\text{tgt}}$ and $s_{\text{tgt2}}$ for CSD. 10 utterances are randomly selected from the LIBRITTS test set to be used for a comparison of baseline, x-vector perturbation, hard cocktail, gradual cocktail, and three-stage cocktail configurations for both the CVC and DUVC systems. This produces a total of 100 unique audio clips for annotation. For 5 randomly selected utterances, each of the 90 crowdworkers listens to 15 samples: 5 ground truth audios, and 5 each from two synthesis configurations applied to the ground truth. Each synthesis sample is thus annotated 9 times, and each ground truth audio clip is annotated 45 times. For each clip, annotators are asked 'How would you rate the naturalness of this speech?', with a scale of bad, poor, fair, good, and excellent [44]. Annotators were required to have at least a 95% approval rating on AMT over 1000 tasks and be located in the United States. At the end of the task, we disclose that some of the speech clips were synthesized.

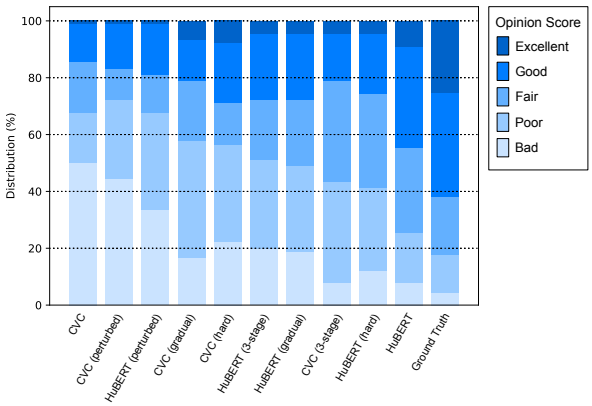



Figure 2: *Opinion score distribution for each voice conversion.*

## 4.4. Results

Results are summarized in Table 1 and Figure 2. Our experiments are informed by the performance of the ground truth nat-

| Model | Naturalness MOS | WER | EER |
|---|---|---|---|
| Ground Truth | 3.65 | 30.8 | - |
| Single speaker: | | | |
| CVC | 1.98 | 39.6 | 21.51 |
| + Perturbation | 2.01 | 31.6 | 25.96 |
| HuBERT | 3.20 | 39.9 | 26.24 |
| + Perturbation | 2.19 | 43.7 | 47.25 |
| Cocktail Decoding: | | | |
| CVC hard | 2.58 | 34.4 | 38.82 |
| CVC gradual | 2.53 | 39.5 | 30.06 |
| CVC 3-stage | 2.74 | 36.7 | 37.11 |
| HuBERT hard | 2.77 | 39.5 | 34.16 |
| HuBERT gradual | 2.64 | 46.6 | 46.79 |
| HuBERT 3-stage | 2.61 | 42.4 | 44.69 |

Table 1: *Results of subjective human evaluation (naturalness), objective word-error-rate (intelligibility), and objective equal-error-rate (speaker verifiability). We aim for high naturalness, low WER, and high EER.*

ural speech samples from the LibriTTS test set. We found that despite the ground truth data being recordings of real human speakers, a non-trivial number of annotators (nearly 40%) rated ground truth speech as "fair" and below. Additionally, the ASR system used for computing WER achieved only 30.8 WER for the ground truth speech, which might be due to the domain mismatch between GigaSpeech and LibriTTS.

Subjective human evaluation determined that ground truth speech sounded the most natural, with a MOS of 3.65. Ground truth speech was evaluated as better ($p < 0.0003$) than the HuBERT voice conversion baseline, with a MOS of 3.20. In turn, HuBERT was more natural than the speaker cocktails ($p < 9 \times 10^{-6}$), which had an average MOS of 2.65. The CVC baseline, as well as all perturbations of baseline models, were the least natural: achieving a low average MOS of 2.06. Figure 2 contains detailed results for human evaluation, with models arranged from least to most natural.

For objective metrics, both CVC and HuBERT baseline voice conversion performed similarly. The introduction of perturbations usually increased both WER and EER, indicating a reduction in intelligibility alongside an increase in anonymization. CSD approaches generally had similar WER reduction, but with further increased anonymization (EER).

### 4.5. Discussion

In comparison to the CVC anonymization baselines, we found that the application of CSD not only dramatically increased the EER (i.e., rendering the speech more anonymous) beyond that of the perturbed x-vector method, it surprisingly improved both the naturalness and the intelligibility of the speech in most cases. We note that simple perturbations also improved the naturalness and WER for CVC, suggesting that local (and continuous) smoothing of the speaker vector improves the perceived quality of the speech. However, despite the perturbation improving the WER of the audio signal, they do not significantly improve the perceived naturalness – which CSD does.

There do however seem to be limitations on the ability of CSD to improve the naturalness of speech. When CSD is applied to HuBERT discrete units, it is detrimental for both naturalness and WER. However, as expected, the EER increases significantly, demonstrating that CSD is a consistently strong strategy for anonymization. Additionally, when compared to x-vector perturbation for HuBERT, CSD produces a similar increase in WER and EER, but maintains much of the naturalness of the speech signal.

The best anonymization strategy appears to be HuBERT with gradual CSD. This strategy provides a dramatic decreased speaker verifiability, while still preserving reasonable naturalness according to subjective human evaluation.

We find that to achieve anonymization, it is not enough to simply convert speech to a single fake speaker; and when perturbing the underlying speaker x-vectors, random noise is not only generally insufficient to disguise the speaker's identity, but damages both intelligibility and naturalness of speech. On the other hand, while CSD can damage intelligibility (though not exceedingly moreso than x-vector perturbation), it maintains the naturalness of speech while at the same time dramatically increasing anonymization.

## 5. Conclusion

In this work, we introduce cocktail speaker decoding (CSD), a novel speaker-shifting strategy which dramatically improves voice-conversion-based anonymization without drastic reduction to speech naturalness. By using CSD, we noticeably encode synthetic speech disclosure in the audio signal, further preventing malicious actors from using publicly released high-quality speech synthesis to deceive humans.

The audio produced by CSD is strange; the constant shifting identity within the waveform produces an "uncanny valley" effect. The speech sounds both natural and *in*human. While we exploit this property to inform listeners that the speech was not spoken by a person, we also imagine that variations of CSD could be used in other speech technologies applications, such as emotional expression or personality design.

## 6. Acknowledgements

The authors of this paper would like to thank Bhiksha Raj, Jeffrey Bigham, Maxine Eskenazi, and Francisco Teixeira for their guidance; and Joshua Zhanson, Han Guo, and other students at CMU School of Computer Science for their helpful comments. J.H. was supported by the NSF Graduate Research Fellowship under Grant Nos. DGE1745016 and DGE2140739. The opinions expressed in this paper do not necessarily reflect those of the funding agencies.

## 7. References

[1] C. Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," *The Wall Street Journal*, vol. 30, no. 08, 2019.

[2] S. A. Buo, "The emerging threats of deepfake attacks and countermeasures," *arXiv preprint arXiv:2012.07989*, 2020.

[3] J. Bateman, *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2020.

[4] E. Wenger, M. Bronckers, C. Cianfarani *et al.*, "" hello, it's me": Deep learning-based speech synthesis attacks in the real world," in *ACM SIGSAC*, 2021, pp. 235–251.

[5] C. Terblanche, P. Harrison, and A. Gully, "Human spoofing detection performance on degraded speech," *Interspeech*, pp. 1738–1742, 2021.

[6] I. Kondratova and B. Emond, "Voice interaction for training: Opportunities, challenges, and recommendations from HCI perspective," in *CHI*. Springer, 2020, pp. 59–75.

[7] "CA Bus & Prof Code § 17941," California State Legislature.

[8] Y. Cho, "Attributable watermarking of speech generative models," Ph.D. dissertation, Arizona State University, 2021.

[9] M. A. Nematollahi and S. A. R. Al-Haddad, "An overview of digital speech watermarking," *International Journal of Speech Technology*, vol. 16, no. 4, pp. 471–488, 2013.

[10] A. Nautsch, A. Jiménez, A. Treiber *et al.*, "Preserving privacy in speaker and speech characterisation," *CSL*, vol. 58, pp. 441–480, 2019.

[11] C.-Y. Li, P.-C. Yuan, and H.-Y. Lee, "What does a network layer hear? analyzing hidden representations of end-to-end asr through speech synthesis," in *ICASSP*. IEEE, 2020, pp. 6434–6438.

[12] Y. Han, S. Li, Y. Cao *et al.*, "System description for voice privacy challenge," *Interspeech*, 2020.

[13] C. O. Mawalim, K. Galajit, J. Karnjana *et al.*, "X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system." in *Interspeech*, 2020, pp. 1703–1707.

[14] B. M. L. Srivastava, M. Maouche, M. Sahidullah *et al.*, "Privacy and utility of x-vector based speaker anonymization," *TASLP*, 2021.

[15] B. M. L. Srivastava, N. Tomashenko, X. Wang *et al.*, "Design choices for x-vector based speaker anonymization," in *Interspeech*, 2020.

[16] F. Fang, X. Wang, J. Yamagishi *et al.*, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 155–160. [Online]. Available: http://dx.doi.org/10.21437/SSW.2019-28

[17] T. Uchida, H. Takahashi, M. Ban *et al.*, "A robot counseling system—what kinds of topics do we prefer to disclose to robots?" in *RO-MAN*. IEEE, 2017, pp. 207–212.

[18] T. Nomura, T. Kanda, T. Suzuki *et al.*, "Do people with social anxiety feel anxious about interacting with a robot?" *Ai & Society*, vol. 35, no. 2, pp. 381–390, 2020.

[19] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," in *Foundations of trusted autonomy*. Springer, Cham, 2018, pp. 135–159.

[20] A. M. Aroyo, F. Rea, G. Sandini *et al.*, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?" *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3701–3708, 2018.

[21] S. M. Shieber, *The Turing test: verbal behavior as the hallmark of intelligence*. Mit Press, 2004.

[22] D. Gros, Y. Li, and Z. Yu, "The RUA-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity," in *ACL*, 2021, pp. 6999–7013.

[23] H. Zen, V. Dang, R. Clark *et al.*, "LibriTTS: A corpus derived from librispeech for text-to-speech," *Interspeech*, pp. 1526–1530, 2019.

[24] B. Sisman, J. Yamagishi, S. King *et al.*, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *TASLP*, 2020.

[25] W.-C. Huang, T. Hayashi, S. Watanabe *et al.*, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading ASR and TTS," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 160–164.

[26] W.-C. Huang, T. Hayashi, X. Li *et al.*, "On prosody modeling for ASR + TTS based voice conversion," *arXiv preprint arXiv:2107.09477*, 2021.

[27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[28] A. Polyak, Y. Adi, J. Copet *et al.*, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[29] J. Shi, X. Chang, T. Hayashi *et al.*, "Discretization and re-synthesis: an alternative method to solve the cocktail party problem," *arXiv preprint arXiv:2112.09382*, 2021.

[30] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *NIPS*, vol. 33, pp. 17 022–17 033, 2020.

[31] V. Panayotov, G. Chen, D. Povey *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[32] S. Watanabe, T. Hori, S. Kim *et al.*, "Hybrid CTC/attention architecture for end-to-end speech recognition," *JSTSP*, vol. 11, no. 8, pp. 1240–1253, 2017.

[33] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

[34] P. Guo, F. Boyer, X. Chang *et al.*, "Recent developments on ESPnet toolkit boosted by conformer," in *ICASSP*. IEEE, 2021, pp. 5874–5878.

[35] Y. Ren, C. Hu, X. Tan *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2020.

[36] D. Snyder, D. Garcia-Romero, G. Sell *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[37] T. Hayashi, R. Yamamoto, K. Inoue *et al.*, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP*. IEEE, 2020, pp. 7654–7658.

[38] T. Hayashi, R. Yamamoto, T. Yoshimura *et al.*, "ESPnet2-TTS: Extending the edge of tts research," *arXiv preprint arXiv:2110.07840*, 2021.

[39] T. Hayashi, "Parallel WaveGAN implementation with Pytorch," 11 2021. [Online]. Available: https://github.com/kan-bayashi/ParallelWaveGAN

[40] N. Tomashenko, B. M. L. Srivastava, X. Wang *et al.*, "The VoicePrivacy 2020 challenge evaluation plan," 2020.

[41] M. McLaren, L. Ferrer, D. Castan *et al.*, "The speakers in the wild (sitw) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[42] G. Chen, S. Chai, G. Wang *et al.*, "Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[43] S. Watanabe, T. Hori, S. Karita *et al.*, "ESPnet: End-to-end speech processing toolkit," *Interspeech*, pp. 2207–2211, 2018.

[44] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *CSL*, vol. 19, no. 1, pp. 55–83, 2005.