



# ClaritySpeech: Dementia Obfuscation in Speech

Dominika Woszczyk<sup>1</sup>, Ranya Aloufi<sup>1,2</sup>, Soteris Demetriou<sup>1</sup>

<sup>1</sup>Imperial College London, UK

<sup>2</sup>Taibah University, Saudi Arabia

d.woszczyk19@imperial.ac.uk, r.aloufi18@imperial.ac.uk, s.demetriou@imperial.ac.uk

## Abstract

Dementia, a neurodegenerative disease, alters speech patterns, creating communication barriers and raising privacy concerns. Current speech technologies, such as automatic speech transcription (ASR), struggle with dementia and atypical speech, further challenging accessibility. This paper presents a novel dementia obfuscation in speech framework, ClaritySpeech, integrating ASR, text obfuscation, and zero-shot text-to-speech (TTS) to correct dementia-affected speech while preserving speaker identity in low-data environments without fine-tuning. Results show a 16% and 10% drop in mean F1 score across various adversarial settings and modalities (audio, text, fusion) for ADReSS and ADReSSo, respectively, maintaining 50% speaker similarity. We also find that our system improves WER (from 0.73 to 0.08 for ADReSS and 0.15 for ADReSSo) and speech quality from 1.65 to  $\sim 2.15$ , enhancing privacy and accessibility.<sup>1,2</sup>

**Index Terms:** Some keywords

## 1. Introduction

As dementia develops, it alters speech patterns, introducing disfluencies, pauses, lexical swaps, and convoluted grammar [1, 2, 3]. These changes not only create communication challenges but also become identifiable markers of the individual's health condition, raising privacy risks and limiting access to speech-operated services [4, 5]. Individuals with dementia are exposed through their voices in both professional and daily life, including identity verification (e.g., voice-authenticated banking) and platforms like podcasts, YouTube, and social media. As speech interfaces like voice assistants and self-driving cars become more common, preserving privacy while maintaining clear communication for dementia patients is increasingly important. Exposed individuals may be flagged as vulnerable, leading to potential discrimination by employers or financial services, or exploitation by malicious agents.

Additionally, despite advances in automatic speech transcription (ASR) [6, 7] current systems face significant limitations in processing dementia-affected and other unusual speech [8]. These systems often struggle to accurately transcribe disordered speech because they are predominantly trained on fluent and clear speech data, resulting in high transcription errors [8, 9]. A naive solution is to resynthesize one's voice to improve its quality but this does not take into consideration the linguistic properties of dementia speech.

Furthermore, existing text-to-speech (TTS) models [10, 11], trained on clean and fluent data, need to be adapted to

atypical patterns which require a significant amount of audio recordings to train on. This is both difficult for dementia patients and creates a possible privacy leakage if the models are trained in the cloud. Finally, restoring and preserving one's original speaker timbre has a crucial impact on an individual's psyche and identity [12].

**Prior Works** Privacy-preserving techniques like anonymization and attribute obfuscation focus on acoustic features and recently approaches such as adversarial techniques and style transfer have been applied, but are challenging with low-resource datasets [13, 14, 15], which is often the case with disordered speech. These systems are also not concerned with transforming the content to preserve privacy, while dementia affects both acoustic and linguistic characteristics of speech. Speech editing systems, such as FluentSpeech [16] or Voicecraft [11], can modify speech locally to correct disfluencies, accents or stutters. Dysarthric speech shares some acoustic features and low data availability with dementia speech. However, an important distinction is that dementia is a cognitive disorder primarily affecting memory, thinking, and social abilities, while dysarthria specifically affects the physical production of speech. Techniques such as data augmentation [17, 18, 19], HMM-based synthesis [20, 21] and voice conversion (VC) [22, 23] have been applied but focus on acoustic features and do not target content.

**Our Approach** In this paper, we propose a framework that utilizes the normalizing nature of zero-shot TTS synthesis and its training-free setting with a text-based obfuscation mechanism aimed specifically at dementia-related patterns. Unlike existing systems that focus on anonymization, noise reduction or disfluency correction, our method obfuscates cognitive impairment patterns while also preserving speaker identity. We evaluate multiple zero-shot TTS models, assessing their ability to minimize dementia leakage (privacy) and the impact of our framework on the ASR utility task, speaker similarity, and speech quality. We outline our contributions below:

- **Holistic Dementia Obfuscation Framework** We propose a novel speech obfuscation framework that tackles cognitive markers in dementia speech while improving its fluency and quality and preserving the speaker's identity.
- **Low-resource Setting** Our system works in a zero-shot setting and does not require fine-tuning.
- **Rigorous Empirical Evaluation** We investigate the privacy (dementia leakage) and utility (ASR, speaker similarity, audio quality) of the proposed framework.
- **Improved Accessibility** We find that our system improves fluency and utility for tasks such as ASR.

<sup>1</sup><https://github.com/domiwk/clarityspeech>

<sup>2</sup>Samples are available at <https://stereomousse.github.io>.

## 2. ClaritySpeech

### 2.1. Threat and Defense Models

We consider an adversary with access to a target user’s recorded speech with the goal of detecting whether the target user suffers from dementia. The adversary can operate in two settings: (a) a static setting and (b) an adaptive setting. A static adversary ( $\mathcal{A}$ ) is obfuscation-oblivious and has direct access to the raw audio. An adaptive adversary ( $\mathcal{A}_{Ada}$ ) is obfuscation-aware. We assume that  $\mathcal{A}_{Ada}$  has gained access to samples generated by any applied obfuscation strategy and can leverage these obfuscated samples to adapt and improve their detection capabilities.

### 2.2. Dementia Obfuscation in Speech

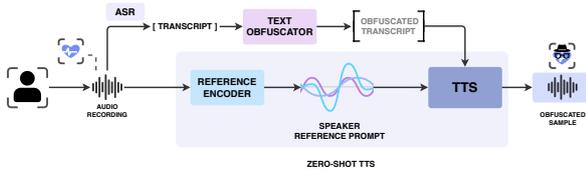


Figure 1: *End-to-end ClaritySpeech framework. The input speech is transcribed through ASR, obfuscated via a text obfuscator and fed to the zero-shot TTS system together with the reference speech, which outputs the obfuscated sample in the reference voice.*

We present an end-to-end framework for dementia obfuscation in speech, ClaritySpeech, consisting of three components: an automatic speech recognition (ASR) module, a text obfuscation mechanism, and a zero-shot text-to-speech (TTS) system. The framework conceals dementia-related attributes while preserving the speaker’s voice and improving intelligibility and naturalness. We describe each component below.

**Automatic Speech Recognition** The first component of our framework is the automatic speech recognition step, which transcribes the audio recording to obfuscate its content. The main challenge of this module is transcription accuracy. Speech affected by dementia can exhibit characteristics such as hesitations, mispronunciations, and non-standard, sometimes incoherent, phrasing making it difficult for ASR systems to make correct predictions. We choose Whisper large v3 [7] as our ASR model, currently the state-of-the-art (SOTA). Whisper has demonstrated robust capabilities in transcribing speech across diverse acoustic environments and speaker conditions. We pick a robust ASR system to minimize the transcription rate even in our atypical setting. The impact of this step on privacy and utility will be analyzed in our experiments section.

**Dementia Obfuscation in Text** Once the content of the speech is transcribed into a text form, the next step involves applying the text obfuscation mechanism. This is the core privacy-preserving step of the framework. The challenges here consist of preserving the original meaning and correcting for incoherence, on top of concealing dementia-specific characteristics and maintaining naturalness. We base our system on DiDOTS [24], which was designed to modify the content of the transcribed text to reduce or eliminate linguistic markers linked with dementia. Through knowledge distillation with a large language model, the model learned to effectively paraphrase sentences and obfuscate dementia while preserving the meaning and intent of the original text. The model has been extensively studied for its

impact on dementia leakage and semantic preservation.

**Zero-Shot Text-to-Speech** The final step in the framework is to synthesize obfuscated speech using a TTS system. We propose to use a zero-shot TTS model that was trained on publicly available data. Current Zero-shot TTS achieve great voice cloning capabilities and can adapt to new voices with minimal data and short reference segments (6s). The reference speech is encoded and used to condition the generation at inference time. This allows us to both not train the model given the low amount of data per speaker but also not to introduce further privacy vulnerabilities. Furthermore, we hypothesize that these models have a normalizing nature. Indeed, zero-shot models work best when trained on very diverse data so they can learn to re-create features seen during training. However, they struggle to reproduce atypical speech (young, old, accents, dysarthric, ...). We want to utilize this “weakness” to normalize dementia speech. We believe that while these models can transfer a general timbre, they will often smooth out granularities such as tempo or unstable pitch and volume. In our experimentation, we compare SOTA models and their impact on the detection abilities of dementia classifiers and acoustic features.

## 3. Experimental Setup

### 3.1. Datasets

We evaluate our systems on the ADReSS (ADR) [25] and ADReSSo (ADRo) [25] datasets. We split our datasets into sentence-level segments. For ADReSS, manual transcripts are provided and we split each sample into segments using the sentence-level timestamps from the transcripts. We split the segments with an 80-20% split for train and test. We further process samples with voice activity detection [26] to trim silences at the beginning and end of segments and remove samples shorter than 3s. We end up with 448 samples (243 CC | 205 AD) in the training set and 242 (140 CC | 102 AD) in the test and validation sets. For ADReSSo, we use the provided segmentation timestamps to isolate segments spoken by the patients, filter them, and get 705 samples in the train set (360 CC | 345 AD) and 242 samples (191 CC | 146 AD) in the test and validation sets.

### 3.2. Adversarial Models

**Models** We select a detection model for each modality, audio and text, as well as their fusion, for our dementia classifiers (adversaries). For the text-based detection model, we re-implement a BERT model based on [27] and [28] with a learning rate of  $1e-6$ , 10 epochs with early stopping (patience=1 on validation loss), gradient clipping of 1, input length limit of 256 tokens, and batch size of 8, trained on augmented data through back translation. For the acoustic-based detection model, we base it on a pre-trained wav2vec model [6], which is a self-supervised model which has shown SOTA performance on the Dementia-Bank datasets [29]. We extract embedding with the wav2vec audio model and feed it to a linear classifier (three layers and a dropout of 0.1) for binary classification. For the fusion of both, we perform early fusion and concatenate BERT and wav2vec embeddings before feeding it to the linear classifier.

**Static and Adaptive Settings** We evaluate our framework in two settings: static and adaptive. Under the static scenario, the adversary has only access to raw data while in the static scenario, the models have knowledge of the obfuscation mechanism used and are trained on both raw and obfuscated samples.

Table 1: Privacy/Utility evaluation of various zero-shot TTS systems on the ADR<sub>ES</sub> and ADR<sub>ES</sub><sub>o</sub> datasets. We report the mean F1 score across static and adaptive adversaries across each modality. The best results for each column are marked in bold. Non-significant results for the static and adaptive adversaries are marked as † and \* respectively.

Dataset	System	Audio↓	Text↓	Fusion↓	Mean Static↓	Mean Adaptive↓	Total Mean↓	SS↑	WER↓	UTMOS↑
ADR <sub>ES</sub>	Original	0.64	0.72	0.73		0.7	0.7	-	0.73	1.65
	ClaritySpeech w/ XTTSv2	0.55†	<b>0.59</b>	<b>0.58</b>	<b>0.49</b>	<b>0.65</b>	<b>0.59</b>	0.50	0.08	2.15
	w/ StyleTTS2	<b>0.54*</b>	0.61	0.62	0.51	0.66	0.61	0.36	<b>0.04</b>	<b>2.86</b>
	w/ Hierspeech++	0.56*	0.64*	0.61	0.52	0.68	0.63	<b>0.52</b>	0.21	2.46
ADR <sub>ES</sub> <sub>o</sub>	Original	0.55	0.67	0.66		0.63	0.63	-	-	1.60
	ClaritySpeech w/ XTTSv2	<b>0.51</b>	0.58*	0.54	0.51	0.58	0.56	<b>0.53</b>	0.15	2.13
	w/ StyleTTS2	0.54*	0.58	<b>0.53</b>	0.54	<b>0.55</b>	<b>0.55</b>	0.39	<b>0.08</b>	<b>2.68</b>
	w/ Hierspeech++	0.55†	<b>0.56</b>	<b>0.53</b>	<b>0.50</b>	0.58	<b>0.55</b>	0.51	0.14	2.62

Table 2: Impact of the different ClaritySpeech components on the privacy and utility metrics of various zero-shot TTS systems, on the ADR<sub>ES</sub> dataset. We report the mean F1 score across static and adaptive adversaries across each modality. The best results for each column are marked in bold. Non-significant results for the static, adaptive adversaries and utility are marked as †, \* and ∇ respectively.

System	Audio↓	Text↓	Fusion↓	Mean Static↓	Mean Adaptive↓	Total Mean↓	SS↑	WER↓	UTMOS↑
ClaritySpeech	0.55	0.59	0.57	0.49	0.65	0.57	0.50	0.08	2.15
w/o ASR	<b>0.47†</b>	<b>0.50*</b>	0.63	<b>0.44</b>	<b>0.63</b>	<b>0.54</b>	0.48∇	0.08∇	2.21
w/o Text obfuscation	0.53†	0.68*	0.66	0.56	0.67	0.62	<b>0.56</b>	0.23	2.15∇
w/o ZS	0.57*	0.60*	<b>0.56†*</b>	0.48	0.65	0.57	0.12	<b>0.02</b>	<b>2.93</b>
w/o ZS & Text obfuscation	0.53†	0.69†*	0.68	0.60	0.68	0.64	0.16	0.29	2.91

### 3.3. TTS Systems

**XTTSv2** was introduced for the task of zero-shot multilingual speech synthesis and achieves SOTA performance in terms of voice cloning [10]. XTTSv2 trains a GPT-based language model for text and audio token predictions conditioned on speaker prompt tokens. The audio codecs are decoded via a Hi-FiGAN [30] and trained with speaker similarity loss. The model is accessed through the coqui-tts package<sup>3</sup>. We pick XTTSv2 as our chosen TTS model given its great cloning abilities but also unlike diffusion and flow-based models, it does not rely on fine-grain prosody control but a reference encoder that learns to transfer the acoustic and stylistic features automatically.

**StyleTTS 2** [31] is a diffusion-based zero-shot voice cloning system trained through adversarial learning which has also achieved SOTA results and naturalness on style transfer and voice cloning. We use the model trained on the LibriTTS train-clean-460 subset<sup>4</sup> and perform inference with the default parameters of  $\alpha = 0.3$ ,  $\beta = 0.7$  and diffusions steps = 5.

**Hierspeech++** [32] is a SOTA model based on a hierarchical VAE, which decouples f0 and durations for more fine-grain control and better style transfer. A denoiser (MP-SENet [33]) is also applied to the samples before extracting features for synthesis and combined with the original style features through a denoising ratio. We use the pretrained checkpoints trained on both English and Korean datasets (2 796 hours)<sup>5</sup> with default parameters and we synthesize the audio samples in 24kHz.

### 3.4. Evaluation Metrics

In our evaluation, we address four research questions: (RQ1) the privacy gain in terms of dementia leakage, (RQ2) the impact of the ASR system, (RQ3) the effect of the zero-shot TTS

system, and (RQ4) the efficiency of each module. To assess **privacy**, we measure the drop in adversarial F1-score for both static and adaptive settings. For **utility**, we compute the word error rate (WER) for ASR, speech quality using the UTMOS score [34], and speaker similarity (SS) through cosine similarity of ECAPA-TDNN embeddings [35]. Finally, we evaluate **efficiency** by measuring latency in seconds and computing the real-time factor (processing time/audio length).

## 4. Results

**Privacy & Utility Evaluation** We evaluate our proposed system for privacy and utility with different zero-shot TTS systems and report the results for the ADR and ADR<sub>o</sub> datasets in Table 1. We find that all systems provide significant privacy and utility gain regardless of the TTS system<sup>6</sup>. However, they exhibit tradeoffs which showcase the importance of selecting the right TTS. On the ADR dataset, XTTSv2 achieves the largest decrease in dementia detection (0.59), outperforming StyleTTS2 (0.61) and Hierspeech++ (0.63). It consistently achieves the best results across all modalities and adversaries (audio static: 0.47; text adaptive: 0.66). StyleTTS2 and Hierspeech++ perform comparably, with StyleTTS2 having slightly better results for static adversaries in audio (0.46 vs. 0.45) but higher leakage overall. On ADR<sub>o</sub>, StyleTTS2 and Hierspeech++ share the best total mean score (0.55), with XTTSv2 slightly behind (0.56). Hierspeech++ performs the best in text and fusion static scores (0.55 and 0.45, respectively), while StyleTTS2 performs better in adaptive fusion (0.57). We note that adaptive adversaries are quite strong, and we hypothesize that given the noisy nature of the original samples, the addition of the cleaner synthesized samples might facilitate learning detailed features. Regarding utility and audio quality, StyleTTS2 achieves the highest utility scores, but it does poorly at speaker

<sup>3</sup><https://github.com/diapa/coqui-ai-TTS>

<sup>4</sup><https://github.com/y14579/StyleTTS2>

<sup>5</sup><https://github.com/shleep/ml/HierSpeechpp>

<sup>6</sup>We perform a McNemar test for the adversaries and Mann-Whitney U for utility scores.

Table 3: Impact of ClaritySpeech on prosody features on ADReSS<sub>GT</sub> (with ground truth transcription), ADReSS<sub>ASR</sub> and ADReSSo test sets. Statistically significant changes are highlighted.

Features	ADReSS <sub>GT</sub>		ADReSS <sub>ASR</sub>		ADReSSo	
	original	ClaritySpeech	ClaritySpeech	ClaritySpeech	original	ClaritySpeech
Syllables num.↑	15.69 ± 8.85	13.17 ± 6.30	11.11 ± 6.67	11.11 ± 6.67	16.22 ± 7.88	11.12 ± 5.14
Energy↑	1.13 × 10 <sup>3</sup> ± 1.15 × 10 <sup>3</sup>	4.99 × 10 <sup>3</sup> ± 7.73 × 10 <sup>2</sup>	4.96 × 10 <sup>3</sup> ± 9.24 × 10 <sup>2</sup>	4.96 × 10 <sup>3</sup> ± 9.24 × 10 <sup>2</sup>	2.43 × 10 <sup>3</sup> ± 1.86 × 10 <sup>3</sup>	4.78 × 10 <sup>3</sup> ± 7.82 × 10 <sup>2</sup>
Pause length↓	438.73 ± 399.55	373.35 ± 174.59	401.98 ± 198.10	401.98 ± 198.10	386.47 ± 206.60	463.87 ± 256.78
F0 <sub>mean</sub> ↑	92.24 ± 44.11	106.40 ± 34.45	100.37 ± 35.18	100.37 ± 35.18	102.37 ± 40.27	94.73 ± 31.90
Pause num.↓	1.66 ± 1.36	0.69 ± 0.91	0.60 ± 0.84	0.60 ± 0.84	1.36 ± 1.28	0.86 ± 0.99
Jitter↓	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.03 ± 0.01
Shimmer↑	0.13 ± 0.05	0.12 ± 0.04	0.13 ± 0.04	0.13 ± 0.04	0.12 ± 0.04	0.12 ± 0.03
Speech rate	2.77 ± 1.08	3.05 ± 0.76	2.73 ± 0.92	2.73 ± 0.92	3.02 ± 0.92	2.64 ± 0.80

similarity (0.36 and 0.39 on ADR, ADRo). Indeed, if the model does not transfer dementia characteristics, its utility and audio quality will improve. Hierspeech++ balances privacy and utility, offering high speaker similarity and moderate privacy leakage across datasets, but also the highest WER (0.21). We find that XTTSv2 strikes the best balance for privacy and utility and speaker characteristics transfer, affirming our design choice.

**Ablation Evaluation** To better understand the impact of each element of our design and ASR, we perform an ablation study and evaluate the sub-systems against privacy and utility metrics. The ASR system seems to have a slight negative impact on the privacy metrics with a mean difference of 3% across all adversaries. However, it has little to no impact on the utility. The text obfuscation module effectively reduces the privacy leakage in Text by 10% and the overall mean by 5%, but reduces speaker similarity, possibly due to the content mismatch. Generating the raw transcribed text from ASR also introduces disfluencies in speech harder to transcribe later on, as indicated by the higher WER. Unsurprisingly, the w/o ZS setting achieves the highest UTMOSS value and lowest WER, but also the lowest speaker similarity as the target voice is unrelated. However, the w/o ZS & Text obfuscation while reaching high audio quality values suffers from poor WER, indicating the importance of the text sanitization step beforehand. We note that the original WER of the ADReSS dataset after transcription is 0.73 (See Table 1), which all of these systems improve on significantly.

**Feature Analysis** Next, we investigate the effect of our framework on the acoustic features of obfuscated samples. We extract features as noted by the literature to be key for identifying dementia in speech and report the mean for the original and obfuscated samples in Table 3. We find that ClaritySpeech significantly modifies prosody features, reducing pauses and increasing pitch (F0), energy, and speech rate. There are however minimal changes in jitter and shimmer. Surprisingly, ClaritySpeech decreases the number of syllables, which could be attributed to the shortening tendencies of the text obfuscation mechanism, making sentences more coherent and concise. The difference between ADReSS<sub>GT</sub> and ADReSS<sub>ASR</sub> does indicate the importance of the content and its impact on the acoustics.

**Efficiency Evaluation** We report the latency of each of the components of our proposed framework in Table 4 on a CPU. We find that the overall system has a non-negligible processing time (average of 11s for samples in the range 3-6s) with a mean real-time factor of 6.07. However, we find that most of this overhead comes from the TTS and ASR modules which on their own introduce a lag of circa 3s and 2s per second of speech respectively. On the other hand, text obfuscation is a relatively lightweight component which is 4x times faster than real-time.

Table 4: Mean and standard deviation of inference time and real-time-factor (RTF) for each ClaritySpeech component across 100 samples on CPU.

System	Mean Time in s (std)	Mean RTF (std)
ClaritySpeech	11.52 (3.94)	6.07 (4.01)
ASR	5.15 (0.74)	3.01 (2.57)
Text obfuscation	0.56 (0.35)	0.25 (0.16)
TTS	6.34 (3.43)	1.71 (0.16)

## 5. Discussion & Limitations

We find that the proposed framework successfully reduces privacy leaks while improving downstream utility, and zero-shot TTS can indeed correct atypicality in this setting. We also find that the impact of the transcription error is mitigated by the text obfuscation mechanism. Limitations of this work include the small, low-diversity datasets used. Future work could combine multiple dementia datasets in English or across languages to improve generalizability. Due to dataset size and quality, we did not fine-tune models in this study. Comparing zero-shot and speaker-dependent models, despite privacy concerns, would be a valuable direction for future research. Finally, future work could also focus on optimizing ASR and TTS for better real-time performance. The proposed framework can be extended to conditions requiring both text and audio modifications (obfuscating speech in recordings for social media, storage, or live obfuscation), but the framework must be tuned for attribute-specific characteristics. Besides privacy, the framework could improve accessibility for individuals with stuttering or disfluent speech, enhancing speech-to-text tools, speech-based services, e-learning platforms, and lecture transcriptions.

## 6. Conclusion

This work addresses the privacy leakage of dementia in speech. We propose an end-to-end framework that transforms both text and audio to conceal dementia markers while maintaining content and speaker similarity. Our approach leverages zero-shot TTS, which filters certain acoustic features and smooths out dementia traits, enhancing downstream tasks. We demonstrate improvements in ASR utility tasks with reduced WER and improved speech quality. Our ablation study also highlights the importance of zero-shot TTS combined with obfuscated text for preserving speaker similarity and fluency. Nevertheless, the privacy and accessibility gains come with the trade-off of diminished speaker similarity. This work lays the foundation for future applications in healthcare and privacy with future work focusing on efficiency and better transfer of speaker identity.

## 7. References

- [1] G. Gosztolya, V. Vincze, L. Tóth, M. Pákási, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.
- [2] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *et al.*, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.
- [3] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.
- [4] B. M. Halpern, T. Tienkamp, W.-C. Huang, L. P. Violeta, T. Rebernik, S. de Visscher, M. Witjes, M. Wieling, D. Abur, and T. Toda, "Quantifying the effect of speech pathology on automatic and human speaker verification," *arXiv preprint arXiv:2406.06208*, 2024.
- [5] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.
- [6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.
- [8] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.
- [9] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking evaluation in asr: Are our models robust enough?" *arXiv preprint arXiv:2010.11745*, 2020.
- [10] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "Xtts: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.
- [11] P. Peng, P.-Y. Huang, D. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," *arXiv preprint arXiv:2403.16973*, 2024.
- [12] J. Mertl, E. Žáčková, and B. Řepová, "Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 4, pp. 342–352, 2018.
- [13] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," *arXiv preprint arXiv:2012.04454*, 2020.
- [14] O. Chouchane, M. Panariello, O. Zari, I. Kerenciler, I. Chihaoui, M. Todisco, and M. Önen, "Differentially private adversarial auto-encoder to protect gender in voice biometrics," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, 2023, pp. 127–132.
- [15] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, pp. 1–14.
- [16] Z. Jiang, Q. Yang, J. Zuo, Z. Ye, R. Huang, Y. Ren, and Z. Zhao, "Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models," *arXiv preprint arXiv:2305.13612*, 2023.
- [17] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *Speech Communication*, vol. 164, p. 103112, 2024.
- [18] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Interspeech 2018*, 2018, pp. 471–475.
- [19] E. Hermann and M. M. Doss, "Few-shot dysarthric speech recognition with text-to-speech data augmentation," in *Proc. INTERSPEECH 2023*, 2023, pp. 156–160.
- [20] C. Veaux, J. Yamagishi, and S. King, "Using hmm-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *Interspeech*, 2012, pp. 967–970.
- [21] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.
- [22] W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, and T. Toda, "A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion," *arXiv preprint arXiv:2106.01415*, 2021.
- [23] D. Wang, S. Liu, L. Sun, X. Wu, X. Liu, and H. Meng, "Learning explicit prosody models and deep speaker embeddings for atypical voice conversion," *arXiv preprint arXiv:2011.01678*, 2020.
- [24] D. Woszczyk and S. Demetriou, "Didots: Knowledge distillation from large-language-models for dementia obfuscation in transcribed speech," *Proceedings on Privacy Enhancing Technologies*, 2025.
- [25] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The address challenge," *arXiv preprint arXiv:2004.06833*, 2020.
- [26] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," <https://github.com/snakers4/silero-vad>, 2024.
- [27] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in *Interspeech*, vol. 2020, 2020, pp. 2162–6.
- [28] D. Woszczyk, A. Hlédiková, A. Akman, S. Demetriou, and B. Schuller, "Data augmentation for dementia detection in spoken language," *Interspeech*. 2022., 2022.
- [29] F. Agbavor and H. Liang, "Artificial intelligence-enabled end-to-end detection and assessment of alzheimer's disease using voice," *Brain sciences*, vol. 13, no. 1, p. 28, 2022.
- [30] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.
- [31] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [32] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," *arXiv preprint arXiv:2311.12454*, 2023.
- [33] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra," *arXiv preprint arXiv:2305.13686*, 2023.
- [34] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.
- [35] B. Desplanques, J. Thienpondt, and K. Demuyne, "Ecapadnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.