Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Voice Reconstruction through Large-Scale TTS Models: Comparing Zero-Shot and Fine-tuning Approaches to Personalise TTS in Assistive Communication

Éva Székely[1], *Péter Mihajlik*[2], *Máté Soma Kádár*[2], *László Tóth*[3]

[1]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
[2]HUN-REN Research Center for Linguistics, Budapest University of Techn. and Econ., Hungary
[3]Institute of Informatics, University of Szeged, Hungary

szekely@kth.se, mihajlik@tmit.bme.hu, tothl@inf.u-szeged.hu

Abstract

Personalised synthetic speech can enhance communication for Augmentative and Alternative Communication (AAC) users, but achieving high-quality, speaker-specific voices depends on various factors such as the condition causing speech loss, and availability of recorded speech. Recent advancements in large-scale zero-shot TTS models may change the data requirements, as they have the potential to adapt to a wider range of inputs. This paper explores the potential of these pretrained models in various data availability scenarios, from extensive spontaneous speech to minimal or no unaffected speech. We evaluate a state-of-the-art TTS system on a case study involving a stroke survivor with dysarthria, leveraging both typical and atypical speech data. Additionally, we introduce a novel interactive approach using dysarthric speech as an audio prompt to enable user-guided prosody adaptation.

Index Terms: speech synthesis, augmentative and alternative communication, dysarthric speech, assistive communication

1. Introduction

Synthetic speech is a fundamental component of Augmentative and Alternative Communication (AAC) systems, supporting individuals with speech impairments to communicate effectively. When synthetic voices represent human speakers, traditional Text-to-Speech (TTS) evaluation metrics take on new significance: "Naturalness" reflects authenticity, "speaker similarity" conveys identity, and "controllability" provides a sense of agency in conversation. AAC research has long emphasised communicative competence over pure intelligibility [1], the importance of interaction over transmission [2], and the role of voice in personal identity [3]. However, despite rapid advancements in speech technologies, individuals who rely on Speech Generating Devices (SGDs) have historically been among the last to benefit from these innovations [4]. The communicative needs of AAC users are highly diverse and influenced by co-occurring conditions such as cognitive or motor impairments. Additionally, speech loss can manifest in different ways – ranging from complete loss to partial retention of speech affected by dysarthria or other atypical features [5, 6]. A key factor in the personalisation of synthetic voices is data availability. While progressive conditions like Amyotrophic Lateral Sclerosis (ALS) or Parkinson's disease may allow time for 'voice banking' [7, 8] – the recording of speech before significant deterioration – other conditions, such as laryngeal tumours, progress rapidly, leaving little time for preparation. And some illnesses – such as stroke – come all of a sudden.

The emergence of a new generation of TTS systems, zero-shot multi-speaker models, trained on vast amounts of speech data, presents significant opportunities for AAC [9, 10, 11, 12, 13]. Fine-tuning these large models for individual speakers can be a way to create higher-quality synthetic voices, especially for those who could not record sufficient amounts of unaffected speech. This raises important questions about the future of voice banking: How much and what type of speech data is truly necessary? How can we leverage non-voice banking recordings, particularly for individuals with sudden-onset conditions? In addition to fine-tuning, the zero-shot capability of these multi-speaker TTS models enables them to generate voices for speakers unseen during training. Many systems claim to reproduce speaker characteristics from as a little amount of data as a few seconds of speech, which could theoretically revolutionise TTS personalisation for individuals with minimal pre-recorded speech or even congenital speech impairments. However, while timbre replication may be feasible with such short samples, the limitations of this approach in capturing supra-segmental features like individual speaking styles manifesting over longer stretches of speech, warrant further investigation to determine its full potential for AAC personalisation.

Open-source neural TTS systems also offer additional advantages, as they are flexible and readily adaptable. This allows for participatory design processes, where speech technology experts, clinical professionals, and end users collaborate to create personalised solutions, echoing Newell's vision of systems that are not only versatile but user-programmable [14]. Such frameworks help combat the limitations of traditional TTS evaluations, and individuals maintain ownership and control over their synthetic voices, and can help ensure that the technology development aligns with their specific communicative needs [15, 16].

The contributions of this paper are: 1) Evaluation of a foundation model-based TTS system across five plausible data availability scenarios for personalised technology-assisted communication. 2) Development and open-sourcing of an online platform that enables an open feedback loop between users and developers, assessed through a video-based survey. 3) Additionally, we propose a new interactional mode, using dysarthric speech as audio prompt in fine-tuned models, and demonstrate its potential for user-guided prosody adaptation.

2. TTS personalisation

2.1. Data availability considerations

We define five plausible levels of speech data availability for personalising TTS systems in AAC. These scenarios depend on the individual's circumstances and the nature of their speech impairment, including whether their condition is progressive, sudden-onset, or congenital.

A) Extensive spontaneous speech available: The individual has access to several hours of spontaneous speech record-

10.21437/Interspeech.2025-1726

ings, which may have been collected prior to the onset of their condition. This could include recordings from professional activities (e.g., lectures, presentations), hobbies, or proactive speech preservation efforts before expected speech decline.

B) Moderate spontaneous speech available: The individual has approximately one hour of spontaneous speech recordings, which may be either available, or feasible to record in the early stages of a progressive condition.

C) Read-aloud sentences via voice banking: The individual proactively records a set of phonetically rich, read-aloud sentences following a diagnosis of a speech-impairing condition.

D) Minimal speech data available: The individual has only a small set of recorded speech samples (e.g., a short message, voicemail, or birthday video) captured before the onset of a sudden speech-impairing event such as stroke or traumatic brain injury (TBI). The available audio may still allow for adaptation through zero-shot TTS methods.

E) No unaffected speech available: The individual lacks premorbid recordings or has a congenital condition resulting in atypical speech patterns, which could potentially be used in speaker-adaptive TTS.

It is important to acknowledge that real-world AAC needs are highly diverse and individualised and can rarely be categorised into the above distinct scenarios. However, for the purpose of systematic evaluation, we define these five levels to create structured comparisons, with the understanding that the insights gained may be adapted to a broader spectrum of user needs.

2.2. Case study with research participant with acquired speech impairment

The speech recordings used in this paper were made available by a native speaker of Hungarian, who, before suffering a stroke, had recorded 13 hours of spontaneous speech in lecture materials (as part of his profession as a university teacher) and 195 read-aloud TTS-tailored sentences as a volunteer in a voice-banking initiative. After the stroke, he developed dysarthria and later recorded the same 195 sentences again with his dysarthric speech, enabling further system evaluation. Following the principles of participatory design, the research participant was actively involved throughout the research process, contributing insights, feedback, and direct experience to shape the development and evaluation of the systems.

This scenario is not commonly encountered, as individuals with sudden-onset conditions such as stroke rarely have both extensive spontaneous speech recordings and TTS-tailored voice banking data recorded prior to their condition. Additionally, studies often lack access to matched pre- and post-stroke speech from the same speaker. The availability of all three datasets in this case allowed us to systematically reconstruct the five hypothetical data availability scenarios outlined above.

We selected XTTS-v2 [9] for our evaluations, because it is an open-sourced multilingual model that includes support for the Hungarian language[1]. Moreover, it is autoregressive, which is an advantage when using spontaneous speech data [17], and it has an effective zero-shot mode capable of replicating speaker characteristics from as little as 10 seconds of audio. In XTTS, zero-shot capabilities are implemented by using a conditioning encoder that extracts speaker characteristics from a short reference audio clip, which is then processed into embeddings that condition the model's GPT-2-based encoder [9]. These embeddings guide the diffusion-based decoder to generate speech that matches the speaker's voice and prosody in multiple languages, even if the model has never seen the speaker during training.

In line with the identified data availability scenarios, the following five systems were used for evaluation:

A) **FT-Spont-Long** Fine-tuned base model on 13 hours of spontaneous speech

B) **FT-Spont-Short** Fine-tuned base model on 1 hour of spontaneous speech

C) **FT-Read** Fine-tuned base model on 195 read-aloud sentences (12 minutes)

D) **ZS-Typical** Zero-shot TTS using the base model with 10s of typical speech recording as reference audio

E) **ZS-Atypical** Zero-shot TTS using the base model with 10s of dysarthric speech recording as reference audio

The spontaneous speech lecture recordings were transcribed with BEAST2 [18, 19] (wav2vec2-uralic fine-tuned on Spontaneous Hungarian speech) and segmented into utterances using breath events as delineator [20]. The fine-tuned models were trained on 1 GPU until validation loss minimised.

3. System evaluations

3.1. Objective evaluation - speaker similarity

For the evaluations, 20 conversational sentences were synthesised with each system (temperature parameter set to 0.65 for all systems, repetition penalty set to 20 for the systems fine-tuned on spontaneous speech, and 10 for **FT-Read** and the two zero-shot systems. Following [9] we compute the Speaker Encoder Cosine Similarity (SECS) between ECAPA2 [21] speaker embeddings extracted from the synthesised samples and held-out audio from the spontaneous corpus. The 20 conversational sentences were compared to 20 held-out samples (See Table 1).

3.2. Perceptual evaluation

To assess the degree to which each of these systems may be suitable to use in the context of assistive communication, we conducted a listening test with a MUSHRA-like setup, 20 sentences in total, 5 systems side-by-side each page. We used situational framing [22]: after a short introduction to the role of synthetic voices in assistive communication, listeners were asked to rate each sample on a 1-5 scale, indicating their opinion on *how suitable it may be as part of a conversation representing human speaker.* 30 native speakers of Hungarian were recruited for the online listening test, which took on average 15 minutes to complete. One-way ANOVA and a post-hoc Tukey multiple comparison test were performed to evaluate the system preferences against each other. This identified that each of the three fine-tuned systems was preferred over each zero-shot system ($p<0.001$), but the difference between ZS-Atypical and ZS-Typical was not significant ($p=0.683$). Between the fine-tuned systems, FT-Spont-Long is preferred

Table 1: *Speaker Encoder Cosine Similarity (SECS) scores*

FT-Read	0.643
FT-Spont-Long	0.727
FT-Spont-Short	0.706
ZS-Atypical	0.324
ZS-Typical	0.635

[1] https://huggingface.co/coqui/XTTS-v2

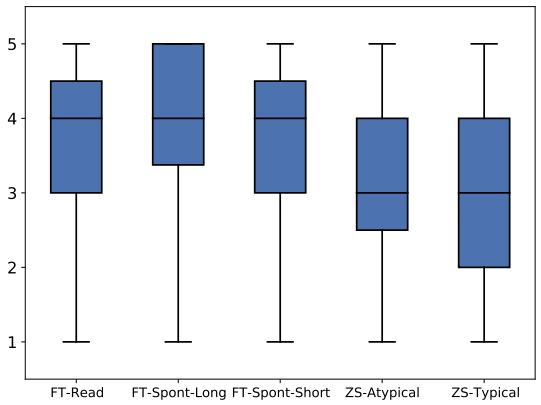

Figure 1: *Results of the MUSHRA-like listening test.*

over FT-Spont-Short ($p<0.001$) and FT-Read ($p=0.0028$). The difference in evaluation between FT-Spont-Short and FT-Read is not significant ($p=0.213$). Samples are available at: https://www.speech.kth.se/tts-demos/interspeech2025-AAC .

4. AAC usability and application

4.1. Web-based interface

We developed an interface using Gradio [23] to facilitate the interactive testing of the various TTS models under development. Users can select a model from a drop-down menu, enter text and optionally provide audio prompts, which can be selected, uploaded, edited, or recorded directly within the interface (See Fig.2). The goal of this platform is to enable a rapid feedback loop between developers, end-users, and clinical professionals, and help ensure that the system personalisation process is guided by all stakeholders in an interactive co-design process. For AAC users who can type on a PC or mobile device, the interface can also function as a standalone communication tool.

4.2. Survey among clinical professionals

To gather feedback on the usability and potential impact of the system, we created an introductory video narrated by FT-Spont-Long, outlining the state-of-the-art in personalised TTS and demonstrating key features of the speech synthesis system and interface. The video was accompanied by a questionnaire and distributed on a forum for clinical professionals in speech therapy and rehabilitation. Respondents were asked to rate: the *intelligibility* of the synthetic voice (1 = non-intelligible, 5 = perfectly intelligible), *naturalness* of the voice (1 = very artificial, 5 = highly natural), *prosody* (e.g., phrasing, emphasis, intonation) (1 = poor, 5 = excellent). Additionally, two open-ended questions invited feedback on expectations for similar systems and any further suggestions.

Ten clinical professionals filled out the survey. All responders were very satisfied with the intelligibility of the synthesised speech as well as with its naturalness, the overwhelming majority of the related answers being 5. Prosody received more mixed feedback, with a majority rating of 4, some respondents noted areas for improvement, particularly in phrasing and emphasis. With regard to the expectations, the most frequent reply was that many neurology patients also have hand mobility issues, making typing not only slow but potentially prohibitive. It coincides with our main concern that to generalise usability we must add

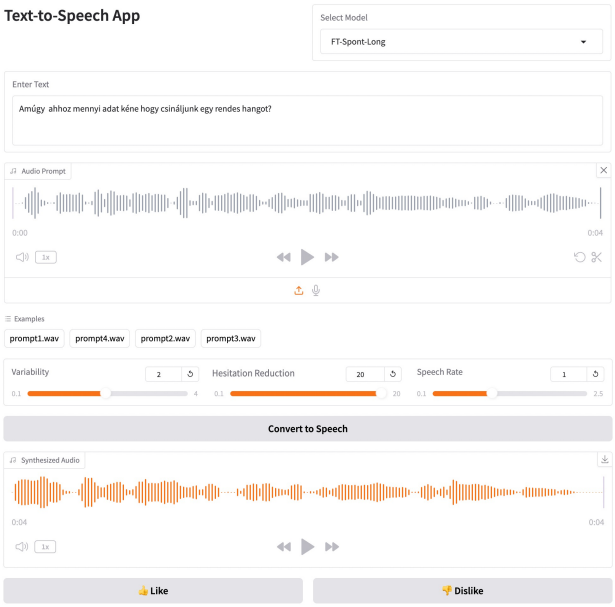

Figure 2: *TTS interface to serve as platform for rapid feedback loop between AAC users, developers and clinical professionals. Users can input text and optionally provide audio prompts, which can be selected, uploaded, or recorded.*

other input modalities. Another relevant remark was that personalisation would indeed be important for the users and that a dysarthric speech input would be particularly beneficial. Lastly, several people said that incorporating context-adaptive speech styles (e.g., casual vs. formal speech) would support practical usability, which is indeed among our future plans.

5. Dysarthric speech as audio prompt

5.1. Speech prompting for prosody adaptation

Reference audios (also called audio/speech prompts) in ZS-TTS do not need to have the same lexical content as the input text. However, as acknowledged in the paper, prosody and speaker information are not entirely disentangled [9], which means that if the lexical content of the audio and text prompt are identical, the system's output will resemble not only the speaker characteristics but also recreate some prosodic features of the input audio. Similar to the zero-shot mode of the base model, fine-tuned systems can also utilise reference audio. This capability opens up the possibility of a new interactive mode in an AAC system, where users not only type their desired message but also record themselves speaking the same sentence with their current (affected) speech. This recording serves as a speech prompt for the TTS system, creating a form of user-directed prosody control. This feature could be particularly beneficial for individuals with aphasia or speech impairments primarily affecting articulation, who retain control over prosodic elements like pitch and emphasis. However, we acknowledge that this approach may not be suitable for all users. To optimise system performance for this interaction mode, we modified the zero-shot algorithm of XTTS to separate the two places where the audio prompt is used. GPT-2 input takes the audio prompt, while the vocoder is fed a vector of the speaker embedding connected to the corpus. To evaluate this, we conducted an experiment measuring

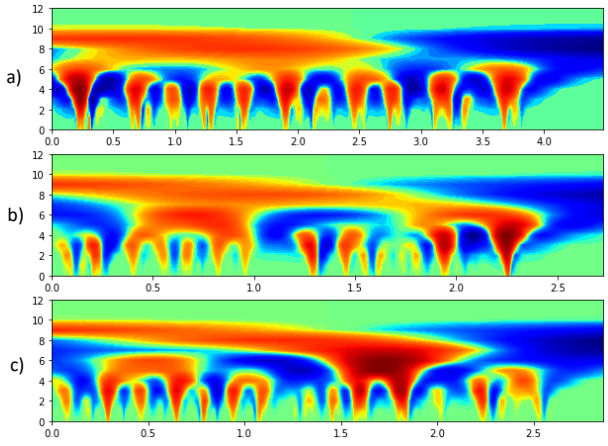

Figure 3: *Scalograms extracted with the Wavelet Posody Toolkit [25]. a) dysarthric speech recording b) the same sentence synthesised with FT-Spont-Long using a) as audio prompt for the GPT-2 in XTTS, and a vector for the speaker embedding c) FT-Spont-Long without audio prompt*

the prosody transfer through prominence peaks, and a qualitative user evaluation of the interactive mode.

5.2. Validating prosody transfer via prominence peaks

The 195 sentences from the voice banking corpus (same as in system **FT-Read**) were recorded again by the speaker, and used to create their dysarthric-prompted synthetic versions, generated with system **FT-Spont-Long**. Another set of synthetic samples were created for comparison, without using a reference audio. CER (Character Error Rate) measured on Whisper-turbo ASR outputs [24] on both prompted and unprompted sets of TTS samples were 2.3% and 2.9%, which shows that the intelligibility was not affected significantly by using dysarthric speech samples in the audio prompts. (For comparison, CER on the dysarthric samples was 68.1%.) To measure whether the output speech resembled prosodic characteristics of the audio prompt, we identified prominence peaks in both sets of synthetic samples and the dysarthric recordings, using the Wavelet Prosody Toolkit [25]. Fig. 3 presents examples of the scalograms of the audio prompt (a) and the two synthetic samples (b) prompted and (c) not prompted. On the normalised length samples, the relative positions of the prominence peaks are comparable. We compare the prosodic characteristics of the two systems to the dysartric speech sample using the Structural Similarity Index (SSIM) between the time aligned scalograms. Using a dysarthric sample aligns prosody better to the speakers intent, demonstrated by an on average 29.3% higher SSIM score (0.0548 vs 0.0424). A binomial test rejects the hypothesis that better outcomes were produced by chance (p=0.0013).

5.3. User Study: Evaluating usability and perceived control

As part of the interactive, qualitative evaluation, the participant conducted 4 conversations, 2 with family members, 1 with a friend, and 1 with a colleague. After each interaction, he documented his observations in a journal. Each conversation partner was also invited to contribute their reflections about their experience. The speaker summarised the main take-aways from the journal in the following points: *1) I noticed that I was able to influence the output of the system with my voice, the speech sounded natural and it gave me a better sense of control over how the speech represented my intentions 2) I didn't think that it was very disturbing to the interaction flow that the sentences were repeated by the TTS, especially since it improved intelligibility, but having to still type in the text slowed down the conversations significantly. 3) Two of the conversation partners noted that this interaction mode might be useful to help new listeners familiarise themselves with the disarthric speech faster, leading to more successful non-mediated interactions.*

6. Discussion

The fine-tuned models (FT-Spont-Long, FT-Spont-Short, and FT-Read) consistently outperformed the zero-shot systems in speaker similarity (SECS scores) and listener preference (MUSHRA test). Among them, FT-Spont-Long performed best, indicating that extensive spontaneous speech data provides the highest-quality speaker adaptation. However, both of the other systems also performed quite well, the gap between FT-Spont-Short and FT-Read was not statistically significant (p=0.213). This suggests that even small datasets can produce highly intelligible and natural-sounding voices, making it a viable alternative when large-scale recordings are unavailable. While the weaker performance of zero-shot settings (ZS-Typical and ZS-Atypical) was expected in both speaker similarity and AAC suitability, a notable finding is that ZS-Atypical performed comparably to ZS-Typical (p=0.683). This demonstrates that zero-shot TTS can adapt to atypical speech as effectively as typical speech, although it may vary slightly with different atypical patterns. This is particularly encouraging for individuals who lack access to typical speech recordings. The results also provide evidence for the viability of such approaches in an under-resourced language like Hungarian [26].

Regarding the interaction mode using dysarthric speech as audio prompt: We conclude that this approach warrants further exploration, as a promising interactive method for prosody adaptation in a text-input AAC system, potentially serving as a viable alternative until voice conversion for dysarthric speech becomes sufficiently advanced to work reliably beyond selected phrases [6] and enable communication without text input.

A limitation of this study is that the interface currently requires typed text input, making it less readily accessible to AAC users with motor impairments. Therefore, it should be regarded as an evaluation platform, rather than a fully functional AAC device. Eye-tracking compatibility will be added in the future.

7. Conclusion

We conclude that large-scale publicly available TTS models are reshaping the landscape of voice reconstruction and personalisation in AAC. Our findings show that even limited amounts of spontaneous speech can be valuable for fine-tuning, offering a promising alternative for individuals who did not have the opportunity to voice bank their unaffected speech. Additionally, the strong performance of zero-shot TTS with atypical speech may enable personalisation for many users with congenital conditions. Finally, our preliminary tests using dysarthric speech as an audio prompt for user-guided prosody adaptation highlight new interactional opportunities in AAC. The open-source interface enables participatory design, as demonstrated in this study, to accelerate the testing of large-scale TTS systems in clinical settings and reduce the time for new technologies to reach those who rely on assistive communication.

8. Acknowledgements

This research is supported by the Swedish Research Council project Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298). The work was also partially supported by the Hungarian NRDI Fund through the projects NKFIH K143075 and K135038, NKFIH-828- 2/2021(MILAB).

9. References

[1] J. Light, "Toward a definition of communicative competence for individuals using augmentative and alternative communication systems," *Journal of AAC*, vol. 5, no. 2, p. 137–144, 1989.

[2] D. J. Higginbotham, K.-E. Kim, and C. Scally, "The effect of the communication output method on augmented interaction," *Augmentative and Alternative Communication*, vol. 23, no. 2, p. 140–153, 2007.

[3] R. Patel and T. Threats, "One's voice: A central component of personal factors in augmentative and alternative communication," *Perspectives of the ASHA Special Interest Groups*, vol. SIG, 12, p. 94–98, 2016.

[4] J. Light, D. McNaughton, D. Beukelman, S. K. Fager, M. Fried-Oken, T. Jakobs, and E. Jakobs, "Challenges and opportunities in augmentative and alternative communication: Research and technology development to enhance communication and participation for individuals with complex communication needs," *Augmentative and Alternative Communication*, vol. 35, no. 1, pp. 1–12, 2019.

[5] D. Liu, Y. Lin, H. Bu, and M. Li, "Two-stage and self-supervised voice conversion for zero-shot dysarthric speech reconstruction," in *2024 International Conference on Asian Language Processing (IALP)*. IEEE, 2024, pp. 423–427.

[6] E. Howarth, G. Vabulas, S. Connolly, D. Green, and S. Smolley, "Developing accessible speech technology with users with dysarthric speech," *Assistive technology: the official journal of RESNA*, pp. 1–8, 2024.

[7] R. Cave and S. Bloch, "Voice banking for people living with motor neurone disease: Views and expectations," *International Journal of Language & Communication Disorders*, vol. 56, no. 1, pp. 116–129, 2021.

[8] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[9] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024, pp. 4978–4982.

[10] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. DAI, H. Lin, J. Chen, X. Du, L. Xue, Y. Chen, Z. Li, L. Xie, Q. Kong, Y. Guo, and W. Xue, "Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis," 2025. [Online]. Available: https://arxiv.org/abs/2502.04128

[11] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[12] Z. Jiang, J. Liu, Y. Ren, J. He, Z. Ye, S. Ji, Q. Yang, C. Zhang, P. Wei, C. Wang *et al.*, "Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[13] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[14] A. Newell, "Today's dream – tomorrow's reality," *Augmentative and Alternative Communication*, vol. 8, no. 2, p. 81–88, 1992.

[15] J. Preece, E. Sullivan, F. Tams-Gray, and G. Pullin, "Making my voice and owning its future," *Medical Humanities*, vol. 50, no. 4, pp. 624–634, 2024.

[16] A. Dietz, S. E. Wallace, and K. Weissling, "Revisiting the role of augmentative and alternative communication in aphasia rehabilitation," *American journal of speech-language pathology*, vol. 29, no. 2, pp. 909–913, 2020.

[17] S. Mehta, H. Lameris, R. Punmiya, J. Beskow, Éva Székely, and G. E. Henter, "Should you use a probabilistic duration model in TTS? Probably! Especially for spontaneous speech," in *Proc. Interspeech*, 2024, pp. 2285–2289.

[18] P. Mihajlik, M. S. Kádár, G. Dobsinszki, Y. Meng, M. Kedalai, J. Linke, T. Fegyó, and K. Mády, "What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced ASR task?" in *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, 2023, pp. 58–62.

[19] M. S. Kádár, G. Dobsinszki, K. Mády, and P. Mihajlik, "Feeding the beast – the enhancement of the BEA speech transcriber and its integration with neural language model (in hungarian)," in *XIX. Hungarian Computational Linguistics Conference*, 2023, pp. 135–143.

[20] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," *Proc. ICASSP*, pp. 6925–6929, 2019.

[21] J. Thienpondt and K. Demuynck, "ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[22] J. Edlund, C. Tånnander, S. Le Maguer, and P. Wagner, "Assessing the impact of contextual framing on subjective TTS quality," in *Proc. Interspeech*, 2024, pp. 1205–1209.

[23] A. Abid, A. Abdalla, D. Botha, S. Guo, Z. Khan, S. Shah, and J. Zou, "Gradio: Hassle-free sharing and testing of [ml] models in the wild," *arXiv preprint arXiv:1906.02569*, 2019.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[25] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[26] É. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, and J. Carson-Berndsen, "Synthesizing expressive speech from amateur audiobook recordings," in *Proc. Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 297–302.