Advancements in Expressive Speech Synthesis: a Review

Shaimaa Alwaisi and Géza Németh

***Abstract*—In recent years, we have witnessed a fast and widespread acceptance of speech synthesis technology in, leading to the transition toward a society characterized by a strong desire to incorporate these applications in their daily lives. We provide a comprehensive survey on the recent advancements in the field of expressive Text-To- Speech systems. Among different methods to represent expressivity, this paper focuses the development of expressive TTS systems, emphasizing the methodologies employed to enhance the quality and expressiveness of synthetic speech, such as style transfer and improving speaker variability. After that, we point out some of the subjective and objective metrics that are used to evaluate the quality of synthesized speech. Finally, we point out the realm of child speech synthesis, a domain that has been neglected for some time. This underscores that the field of research in children's speech synthesis is still wide open for exploration and development. Overall, this paper presents a comprehensive overview of historical and contemporary trends and future directions in speech synthesis research.**

***Index Terms*—Speech style, Expressivity, Emotional speech, Expressive TTS, Prosody modification, Multi- lingual and multi-speaker TTS**

I. INTRODUCTION

THE objective of this study is to explore the latest advancements in speech synthesis research. It is primarily intended for researchers involved in the development and enhancement of Text-to-Speech (TTS) systems, as well as professionals in various fields that utilize TTS applications, including such as customer service, navigation systems, and language education [1].

TTS is a process that converts written text into speech like that of humans [2]. Speech serves is a crucial element in human interaction and verbal communication. Throughout history, people have relied on speech as an effective means of conveying information, expressing themselves, and revealing their emotional state [3]. We communicate using various speech styles, which can differ based on factors such as the subject, environment, and culture [2]. In other words, speech styles depend on the content, context, and audience. They can range from formal to casual.

In recent years, advancements in speech technology have led to the development of artificial speech that closely resembles human speech in terms of naturalness and intelligibility. This technology, also known as speech synthesis, takes text as input and generates speech as output. Modern TTS systems have evolved from a long history of efforts to create synthesized human language from written text.

Numerous TTS applications have achieved impressive levels of naturalness and intelligibility. Key factors contributing to naturalness include expressiveness, emotion, and speech style. Modern TTS systems need to deliver synthesized speech in the desired style for users. Expressivity pertains to the manner in which thoughts, emotions, and information are conveyed through a specific expressive style [1] [4].

Speech style in speech synthesis is influenced by various factors, such as the topic, language, speech rate and intensity, and regional culture of the spoken language. Linguistically, expressivity refers to communicating positive or negative ideas or emotions in a style that is relevant to the listener. Emotional expression serves as a vocal indicator of emotions, which is evident in the speech waveform [5]. In addition to speech styles, emotions are also considered expressions. Different expressive styles can be generated based on two approaches: corpus-driven and prosodic-phonology approaches. The corpus-driven approach involves analyzing large datasets of speech to extract patterns of prosody associated with different emotions. This data is then used to train a machine learning model to predict the appropriate prosody for a given text.

The prosodic-phonology approach, on the other hand, involves modeling the underlying linguistic and phonological features of speech that give rise to different emotional expressions. This approach involves analyzing the sound units such as fundamental frequency F0 and duration [6].

Department of Telecommunications and Media Informatics Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary
(E-mail: shaima.alwaisi@edu.bme.hu, nemeth@tmit.bme.hu)

DOI: 10.36244/ICJ.2024.1.5

Neutral Speech Synthesis Systems (NSS) generate speech from written text in a single style, often referred to as neutral or flat sound [7]. Fig 1. presents a framework for a Neutral Speech Synthesis System (NSS). The text analysis and processing stage known as front-end involves the analysis of the written text to extracting linguistic features. This stage provides linguistic and acoustic features to the back-end stage, where the acoustic features of the speech signal are generated.

The back-end stage is where the speech signal is synthesized from the linguistic features extracted in the text analysis and processing stage. This stage involves converting the linguistic features into acoustic features, such as pitch, duration, intensity, and spectral characteristics, to generate a natural-sounding speech signal [8]. Linguistic features are derived through syntactic, semantic, and lexical analysis steps, which guide the synthesis process to produce neutral speech [9]. To generate speech with a specific expressive style, the desired expressive style is incorporated as an additional input to the TTS model, as depicted in Fig. 1 [10] [11].

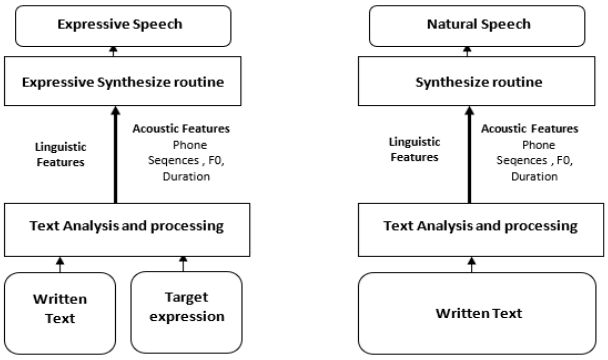


Fig 1: TTS and Expressive TTS System Architecture. On the left side, a schematic diagram illustrates the expressive TTS system. This system processes input text along with the desired expressive elements. On the right side, natural speech is generated by a natural TTS system [12].

The benchmark for evaluating speech technologies is human ratings. Traditionally, listeners are tasked with listening to speech samples and providing ratings, either in isolation or within a context. However, researchers face challenges in evaluation of the new system, since Ratings are subjective, varying from person to person. This subjectivity becomes more pronounced when listeners have limited context and training [13].

The most common method of evaluation is the MOS (Mean Opinion Score) test [14]. This test involves collecting MOS scores from listeners who evaluate each utterance in isolation. In this method, listeners assign scores to individual utterances, typically on a five-point scale, with 5 score representing highly natural speech and 1 score representing highly unnatural speech. Unlike MOS tests, where ratings are provided in isolation, the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test involve listeners in a multiple comparison test. MUSHRA test offers enhanced the sensitivity to subtle differences between stimuli compared to MOS tests [13].

This paper follows the following structure: Section II, since current TTS systems based on the advance Deep learning algorithms, we first introduce several Deep learning models widely used in TTS systems. In Section III and IV, we review some of subjective and objective metrics that are used for evaluating TTS models. In Section V, we summarize style representation and transfer methods. Section VI. discusses prosody modelling in speech synthesis. In Section VII. we take attempt to point out some challenges in child speech synthesis. Finally, in Section VIII, the paper concludes with a summary of the findings and possible future directions.

II. DEEP LEARNING BASED SPEECH SYNTHESIS

Deep Neural Networks (DNNs) [15] play a crucial role in modern speech synthesis approaches, such as WaveNet [16] and Tacotron [17]. The shift from decision trees to deep learning methods has led to significant improvements in the quality of synthesized speech. This shift also involves a move from (HMM) to frame prediction using deep learning models, contributing to the notable improvements in speech synthesis quality. However, a closer examination of the literature reveals several challenges, including the need for substantial computational resources and large speech datasets to train TTS models. Moreover, recording speech datasets with professional speakers can be costly. These challenges have been addressed through various knowledge transfer approaches, such as fine-tuning, transfer learning, and multi-task learning [18].

A. Back-End Synthesizer

WaveNet, proposed by Google DeepMind in 2016, is a deep learning-based autoregressive approach. This fully probabilistic and autoregressive model generates synthesized speech that closely resembles natural audio waveforms. WaveNet's architecture is based on a Convolutional Neural Network (CNN) trained with speech samples to predict natural speech, with each sample depending on the previously generated ones. WaveNet serves as a vocoder for TTS models, with inputs consisting of linguistic features, predicted log fundamental frequency (F0), and phoneme durations [18].

For expressiveness prediction, non-autoregressive WaveNet blocks outperform the original WaveNet [19]. Multi-speaker WaveNet vocoders have demonstrated higher performance compared to traditional methods [20]. Parallel WaveNet combines WaveNet and the Inverse Autoregressive Flow (IAF) method. Inverse-autoregressive flows (IAFs) are generative models used for high dimensional observable samples. High-dimensional observable samples typically refer to a large set of acoustic features that are extracted from speech signals, such as the fundamental frequency, spectral envelope, and time-varying spectral parameters [21]. Capable of generating speech in a wide range of styles (emotional, neutral, conversational, long-form reading, news briefing, and singing), Parallel WaveNet [22] uses a vocoder trained on a multi-speaker emotional dataset to convert Mel spectrograms from a neutral style to various emotional styles [23].

A novel speech synthesis system called Autovocoder has been proposed by [24] to generates high-quality audio and outperforms other waveform generation systems in terms of computational cost. Autovocoder is trained as a denoising autoencoder and generates a waveform at a speed 5 times greater than Griffin-Lim algorithm [25] and 14 times faster than the neural vocoder HiFi-GAN [26].
Autovocoder utilized parallel computing and data parallelism techniques by leveraging fast, Differentiable Digital Signal Processing DSP operations, a purely convolutional residual network, and a learned representation to achieve efficient and fast waveform generation.
Generative Adversarial Networks (GANs) have emerged as a powerful tool for generating high-quality audio, including speech synthesis. GAN-based vocoders are a type of vocoder that uses GANs to generate raw waveform audio from acoustic features and linguistic information. This approach offers several advantages over traditional vocoders, such as improved audio quality, expressiveness, and robustness to noise. Among the various GAN-based vocoders that have been developed, prominent instances of well-known models include HiFi-GAN [26], SnakeGAN [27],Parallel WaveGAN [28], and BigVGAN [29].

B. Linguistic Analysis and Prosody Front-End

Front-end models play a crucial role in processing input text into intermediate representations, often involving linguistic features or phonetic information. Tacotron is an end-to-end Text-to-Speech (TTS) system that uses deep neural networks to generate natural-sounding speech from text input. It operates by predicting mel-spectrograms from text characters, which are then converted into time-domain waveforms using a vocoder [30]. Tacotron2 is a generative model that combines an encoder-decoder architecture with a soft attention mechanism to generate spectrograms from a given text [31]. The primary concept of the attention mechanism is to identify the most relevant characters for each Mel spectrogram frame and determine weights for each character embedding [31].

Tacotron2 has been employed to enhance the expressivity of multi-speaker end-to-end TTS models. The expressivity of latent representation is used for predictions made by the encoder to derive emotion [32]. Text-Predicting Global Style Token (TP-GST) is combined with Tacotron to generate speech in a specific style. Style attention, prosody encoder, and style embedding are added to Tacotron. During the training phase, a combination of trainable embeddings is extracted to be shared across the entire text, driving the Global Style Tokens [33].

FastSpeech [34], another notable front-end model, introduces a novel feed-forward network that generates mel-spectrograms in parallel, utilizing feed-forward Transformer blocks, a length regulator, and a duration predictor. FastSpeech incorporates a phoneme duration predictor to ensure hard alignments between phonemes and mel-spectrograms, reducing the ratio of skipped words and repeated words and contributing to high audio quality. FastSpeech aims to address several challenges present in traditional autoregressive text-to-speech (TTS) models. These challenges include slow inference speed, lack of robustness leading to word skipping and repeating, and limited controllability over voice speed and prosody.

TABLE 1
WAVENET-BASED EXPRESSIVE VOCODERS

Reference	Expressions	Evaluation method	Parameters	Findings
[23]	Happy, angry, sad	Mean Opinion Score (MOS)	linear-scale log magnitude spectrograms and mel spectrograms Using dynamic time warping (DTW) to align them	Implementing WaveNet vocoder to generate speech from melspectrograms led to overall improvement regarding the quality of synthesized speech of each emotion.
[98]	Normal, happy, angry	Mean Opinion Score (MOS)	mel-spectrum parameters and Emotion ID (EID)	Proposed model successfully gener emotional speech t into account mel-spectrum parameters
[22]	Emotional, Neutral, Conversational Long-form reading, News briefing and Singing	MUSHRA	mel-spectrograms	The proposed method synthesized speech with various styles and languages in real-time

FastPitch [35], a fully parallel text-to-speech model, draws its foundation from Fast Speech. FastPitch conditioned on fundamental frequency contours. It predicts pitch contours during inference, allowing for more expressive and engaging speech. The model retains the favorable, fully parallel Transformer architecture.

On the other hand, FastSpeech2 [36] represents a paradigm shift in text-to-speech modeling as a non-autoregressive system. FastSpeech2 simplifies the training pipeline, improves voice quality, and introduces more variation information of speech, such as pitch and energy, as conditional inputs. It provides variance information such as pitch, energy, and more accurate duration. The model architecture includes a pitch predictor, pitch contour, and pitch spectrogram, allowing for manual manipulation of pitch, duration, and energy in synthesized speech.

III. SUBJECTIVE METRICS

1- Mean Opinion Score (MOS): This is a widely used subjective measurement for evaluating the quality of synthesized speech. Listeners rate the speech using a numerical scale ranging from 1 to 5, where 5 signifies excellent speech quality and 1 represents the lowest quality. MOS is a subjective method recommended by standardization bodies such as IEEE Subcommittee. During the listening test, listeners complete a questionnaire that may include sections on overall impression, listening effort, pronunciation, speaking rate, articulation, and voice pleasantness [37].

2- AB Preference Test: In AB Preference Test, participants are presented with audio samples from two distinct speech synthesis models, denoted as model A and model B. Participants listen to samples from both systems and express their preference [38].

3- ABX Preference Test: Participants listen to three speech versions—A, B, and X—with X being the target speech and A and B being two synthesized speech sentences generated by different models. Test subjects are asked to choose which synthesized version is closer to the target speech X[38].

4- MUSHRA (Multiple Stimuli with Hidden Reference and Anchor): In a MUSHRA test, participants evaluate systems on a scale ranging from 1 to 100. They accomplish this by listening to stimuli for the same text presented side-by-side, in comparison to a high-quality reference. This method facilitates a comprehensive assessment of multiple systems, allowing for a nuanced ranking based on perceived quality [39].

IV. OBJECTIVE METRICS

Objective measurements involve the quantitative evaluation of speech synthesis systems, providing a mathematical assessment of the quality of synthesized speech.

A. Itakura-Saito measure

This method is like most objective methods for the evaluation of TTS Models divides the speech signal into frames. Let $s(i)$ and $s'(i)$ be two sampled speech signals, and $x_n(i)$ and $x'_n(i)$ are two windowed frames generated from implementing a window equation $w(i)$, where n is the frame index designating the window location.

$$x_n(i) = w(i)s(i + n) \quad (1)$$

$$c = w(i)s'(i + n) \quad (2)$$

We indicate the z-transform of $x_n(i)\ and\ x_n(i)$ by $X_n(z)$ and $X'_n(z)$. The Fourier transform is derived by assessing the z-transform on the unit circle, i.e., $z = e^{j\omega}$. The $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$ are utilized to represent the Fourier transforms of two signals that have been windowed, respectively. Then for each pair of $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$, spectral distortion $p[Xn, X']$ is defiend as the dissimilarity among $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$, the Itakura-Saito formula for speech analysis is defined as below [40].

$$p_{is}[X_n, X'_n] \triangleq \int_{-\pi}^{\pi}\left[\left(\frac{|X_n(e^{j\omega})|^2}{|X'_n(e^{j\omega})|^2}\right) - \Lambda(\omega) - 1\right]\frac{d\omega}{2\pi} \quad (3)$$

where

$$\Lambda(\omega) = log\ |X_n(e^{j\omega})|^2 - log\ |X'_n(e^{j\omega})|^2 \quad (4)$$

B. *Root mean square (RMSE)*

RMSE is a mathematical measure used to evaluate log f_0 trajectories produced by TTS models, it is stated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(F0_i) - \log(F0'_i))^2} \quad (5)$$

Where $F0_i$ and $F0'_i$ stand for the original and predicted F0 features, respectively. and N is the length of the F0 sequence [41].

C. Gross pitch error (GPE)

GPE is the proportion of segments that are measured as voiced for natural and generated speech having relative pitch error higher than a certain threshold (usually taken as 20% in speech analysis) [42].

V. STYLE REPRESENTATION AND TRANSFER

A. Global Style Token

In text-to-speech, Global Style Tokens (GSTs) are a recently proposed method for extracting style embedding features that reflect specific speech styles. GSTs introduce an auxiliary input vector to the speech synthesis model to control the global style of the synthesized speech. Style tokens are global features of speech style that can be adjusted to synthesize speech in a target style. Modern GST architectures have been developed to learn latent representations of high-dimensional speech data [43] . An attention mechanism calculates attention weights for style tokens, and the sum of style tokens is used for style embeddings. During the training phase, style tokens are initially created randomly, and they learn speech styles in an unsupervised manner.

In [44], a Global Style Token (GST) network is combined with an augmented version of Tacotron to capture expressive variations in speech style. The GST network processes GST combination style embeddings as expressive style labels that are jointly predicted within Tacotron. The TP-GST network extracts weights or style embedding space from text alone, without explicit labels during training phases. Two text-prediction pathways, Predicting Combination Weights (TPCW) and Predicting Style Embeddings (TPSE), are used to extract style tokens during inference time. TP-GST methods successfully generate expressive speech without background noise. Other studies [45] [46] [47] [48] have also utilized GSTs in various ways to synthesize expressive speech, control speaking styles, and explore fine-grained control of speech generation.

Inspired by the GST module, [49] proposes using global speaker embeddings (GSEs) to control the style of synthesized speech. GSE has a unique purpose and functionality that differs from GST such as focusing on capturing the speaker-specific characteristics within a given text, enabling the identification of speakers from their speech patterns. In contrast, GSTs are designed to capture the stylistic elements of a text, such as reading or formal styles. They enable the modification of text style while preserving its content.

In general, GSTs are an effective method for controlling global stylistic features of synthesized speech. However, they have limitations and challenges, such as requiring a sufficient amount of speech samples during the training step to effectively synthesize speech in the desired style. As GSTs are designed to capture global style features, they may not be an effective tool for controlling the nuances of the desired style, such as intonation or rhythm.

B. Style disentanglement

Speech style disentanglement refers to the process of extracting various style factors, such as prosody, speaker, and linguistic-related factors, which enables fine-grained control of multi-reference speech style on separate speech datasets. Disentangling different informative factors in speech synthesis is essential for highly controllable speech style transfer. One of the significant challenges in speech technology is separating intertwined informative factors. Therefore, separating representations of these factors can enhance the robustness of expressive speech synthesis systems [50]. Traditional latent space representation learning algorithms predict general style embeddings with limited fine-grained control.

In [51], disentangled latent space representations based on adversarial learning are adopted to improve the robustness of highly controllable style transfer in voice conversion (VC). An Adversarial Mask-And-Predict (MAP) network is designed to explicitly disentangle the extracted speech representations, which include content, timbre, and two additional factors related to prosody, rhythm, and pitch. (MAP) network consists of a gradient reverse layer (GRL) and a stack of prediction head layers. During training, one of the four speech representations is randomly masked, and the adversarial network attempts to infer the masked representation from the other three representations. The prediction head layers in the MAP module are composed of a fully connected layer, GeLU activation, layer normalization, and another fully connected layer. The MAP network is trained to predict the masked representation as accurately as possible by minimizing the adversarial loss. However, during backward propagation, the gradient is reversed, which encourages the representations learned by the encoder to contain as little mutual information as possible.

The adversarial MAP network aims to increase the correlation between the masked and other speech representations, while the speech representation encoders try to disentangle the representations to decrease the correlation using the inversed gradient of the adversarial MAP network. The proposed method enhances the quality of synthesized speech in voice conversion across multiple factors [52]. A single model is trained for multiple speakers using the adversarial learning framework, instead of building a separate model for each target speaker. The proposed method has two training phases, resulting in significant improvements in the quality of synthesized voice. In [53], a zero-shot style transfer approach using disentangled speech representation learning is adopted to transfer speech styles with non-parallel datasets. The disentanglement process improves style transfer accuracy.

In general, disentanglement speech representation learning is a promising approach for highly controllable speech style transfer. However, this method comes with computational complexity that requires substantial computing power. This issue needs to be carefully considered.

C. Cross-speaker style transfer

Cross-speaker style transfer (CSST) is a cutting-edge technique for synthesizing expressive speech. It aims to transfer multiple speaking styles from various supporting speakers to a target speaker while maintaining the target speaker's identity and timbre [54]. Unlike traditional speaking style transfer methods that collect style embeddings from reference speech and use them as auxiliary inputs to synthesize stylized speech [55], [56], modern cross-speaker style transfer conveys different speaking styles between speakers without requiring text-paired reference speech [57]. Numerous studies have adopted CSST to transfer speech styles between multiple speakers.

In [58], a chunk-wise multi-scale cross-speaker style and adversarial classifiers are proposed for style transfer. Multi-scale cross-speaker style is trained in two phases to predict both global style embeddings (GSE) and local prosody embeddings using an adversarial training approach. An adequate amount of speech style data from non-target speakers is needed during the training process. In [59], a multi-speaker acoustic system called Daft-Expert is employed to transfer highly expressive prosodic styles from both seen and unseen speakers. FiLM conditioning layers are used to embed prosody information in the TTS system. FiLM conditioning layers is a general-purpose conditioning technique for neural networks known as FiLM (Feature-wise Linear Modulation) proposed by [60]. FiLM layers influence neural network computations through a straightforward feature-wise affine transformation, utilizing conditioning information. The proposed model is combined with both FiLM layers and adversarial learning for highly accurate cross-speaker transfer.

Cross-speaker transfer with data augmentation techniques has been successfully used in low-resource expressive TTS systems. A recent study [61] applied data voice conversion VC-based augmentation for cross-speaker style transfer, where expressive speech datasets are not available for the target speaker. The adopted method uses two models: Pitch-Shift PS-based data augmentation and voice conversion VC-based data augmentation. Pitch-shift PS-based augmentation involves altering the fundamental frequency of the speech signal, providing a technique to modify the perceived pitch without changing the speaker identity .PS-based augmentation is used for source and target speaker samples to enhance the stability of the training stage, while short-time Fourier transform (STFT)-based optimization is adopted for the voice conversion training stage.

FastSpeech Multi-language TTS system [62] applied cross-language style transfer to synthesize speech in any speaker style in the target language, overcoming the challenge of non-authentic accent issues in cross-speaker style transfer. Conditional variational encoder and adversarial learning are used in the training process. Cross-speaker style transfer still faces challenges since multiple speakers have varying styles and timbres. Several studies have applied different techniques, such as speaker normalization [63] [64] [65] to model speaker attributes, data augmentation [66] [67] [68] [69] , and multi-task learning [70] [71], to generalize TTS systems to new speakers.

D. Speaker adaptation

TTS systems that employ speaker adaptation techniques aim to adjust a pre-trained model with a large-scale corpus to accommodate unseen speakers during the training process, even when there is a limited amount of speech data. Speaker adaptation is an effective technique when only a few minutes of target style data are available, as its primary role is to transfer speaking styles from a source speaker to a new speaker with limited adaptation data [72]. Adaptation strategies can be divided into two main categories. The first category of TTS systems uses pre-trained additional encoding networks to predict speaker attributes, which are then combined with linguistic characteristics as inputs to the synthesizer model [73] [74] [75] [76]. On the other hand, the second category fine-tunes the weights of the pre-trained multi-speaker TTS system to mimic a new speaker [77] [78]. Bayesian optimization (BO) has achieved high performance in fine-tuning TTS models.

A novel method called BOFFIN TTS (Bayesian Optimization for Fine-tuning Neural TTS) has been able to transfer styles for voice cloning in TTS systems under data-scarcity constraints [79]. This proposed method finds the optimal weights for hyperparameters for any target speaker in a functional and automatic manner. One of the critical aspects of this approach is its ability to intelligently search the hyperparameter space while minimizing the required computational resources. This is achieved through the use of Gaussian processes, which model the target function and provide a measure of uncertainty to guide the search for optimal hyperparameters. By exploiting this uncertainty, the algorithm can effectively balance exploration and exploitation during optimization. Another advantage of the Bayesian optimization approach is its flexibility in incorporating various constraints and domain knowledge into the optimization process. For example, one can introduce regularization terms or prior information on the hyperparameters to improve the adaptation performance. This can be particularly useful when dealing with challenging scenarios, such as limited data or highly diverse speaker characteristics.

Some recent works have also explored the combination of Bayesian optimization with other machine learning techniques, such as transfer learning and multi-task learning, to further improve the adaptation process [80]. By leveraging the shared information between different speakers or tasks, these approaches can achieve better performance, even with limited adaptation data. Despite the promising results, there are still challenges in applying Bayesian optimization for speaker adaptation in TTS systems. One of the main issues is the scalability of the optimization process, as the complexity of Gaussian process regression grows with the number of observations. This can limit the applicability of the method to large-scale problems or high-dimensional hyperparameter spaces. Moreover, the choice of the surrogate model and acquisition function, as well as the initialization of the optimization process, can significantly impact the overall performance.

VI. PROSODY MODELLING IN SPEECH SYNTHESIS

Prosody is a crucial aspect of speech synthesis that focuses on the rhythmic, melodic, and expressive features of speech [76]. The primary components of prosody include pitch, duration, intensity, and pauses, which collectively contribute to the overall expressiveness and naturalness of synthetic speech. Prosody helps convey emotions, emphasis, and linguistic structure in spoken language, thus playing a significant role in making synthetic speech sound more natural [77]. Hidden Markov Models (HMMs) have been used for capturing prosodic and linguistic features of speech, where decision trees are used to tie contextual features to individual nodes of the decision tree [81]. This approach enables more accurate modeling of prosody, allowing for generating natural and expressive synthetic speech. A new approach has been applied for prosody modeling [82]. This approach enhances prosody by integrating pre-trained cross- utterance (CU) representations from Wav2Vec2.0 and BERT into Fastspeech2. It improves speech naturalness and expressiveness in Mandarin and English but heavily relies on pre-trained models and lacks evaluation on other languages. Further investigation into model layers is needed for better prosody modeling.

A. Pitch Contour Modeling

Pitch contour modeling is the process of estimating and generating the fundamental frequency (F0) of speech, which corresponds to the perceived pitch. Accurate pitch contour modeling is essential for achieving natural-sounding prosody in speech synthesis. Many studies have been conducted to enhance the robustness of pitch. Among them, FastPitch [35] has gained popularity for its ability to control pitch and duration at the phoneme level during the synthesis of speech by conditioning these values. VocGAN-PS [83] and the FastPitch training algorithms have been proposed to improving pitch controllability. VocGAN-PS is a timbre-preserving pitch shift method that expands the pitch range without altering vocal characteristics. It avoids the need for additional algorithms like pitch tracking, however, may struggle with precise pitch estimation during transitions. The FastPitch training algorithm utilizes pitch-augmented speech data generated by VocGAN-PS to enhance FastPitch's pitch control and robustness, but its effectiveness relies on the quality and diversity of the augmented datasets.

There are different techniques for pitch contour modeling, including rule-based methods, statistical parametric methods [84], and deep learning approaches [81]. Rule-based methods use linguistic and phonetic rules to generate pitch contours, while statistical parametric methods (e.g., hidden Markov models or Gaussian mixture models) learn the relationship between linguistic features and pitch contours from data. Recently, deep learning methods like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have also been employed for pitch contour modeling, leveraging their ability to learn complex patterns and capture long-range dependencies in the data [67].

B. Duration Modeling

Duration modeling deals with predicting the duration of phonemes, syllables, or words in synthetic speech. Accurate duration modeling is vital for natural-sounding speech, as it contributes to the overall rhythm and pace of the spoken language [85].

Reference [86] propose an unsupervised text-to-speech (UTTS) system. In this system, a Speaker-Aware Duration Prediction module takes the phoneme sequence and speaker embedding as input to predict the speaker-aware duration for each phoneme. The phoneme sequence is first passed into a trainable look-up table to obtain the phoneme embeddings. Then, a multi-layer attention module is used to extract the latent phoneme representation, followed by a conv-1D module to combine the latent phoneme representation with the speaker embedding. A linear layer is then applied to generate the predicted duration in the logarithmic domain. During training, the Mean Squared Error (MSE) is utilized to calculate the difference between the predicted duration and the target duration obtained from forced alignment extracted by Montreal Forced Alignment (MFA).

During inference, the duration predictor rounds up the predicted duration and expands the phoneme sequence to form an estimated forced alignment. This estimated forced alignment is then used in the UTTS system for speech synthesis.

In [87] zero-shot TTS model utilized duration modeling as part of the conditioning process, enabling rhythm transfer and extracts disentangled embeddings between rhythm-based

speaker characteristics and acoustic-feature-based ones.The proposed method captures rhythm-based speaker characteristics , leading to higher perceived speaker similarity.

Another study [88] proposed two approaches to improve duration modeling in TTS systems. The first approach is a duration model conditioned on phrasing, which enhances predicted durations and provides better modeling of pauses. The second approach is a multi-speaker duration model called Cauliflow, which utilizes normalizing flows to predict durations that better match the target duration distribution. The proposed models improved naturalness of speech and variable durations for the same prompt, as well as variable levels of expressiveness.

C. Intensity Modeling

Intensity modeling is concerned with estimating and generating the energy or intensity of speech signals. Intensity contributes to the perceived loudness and stress patterns of synthetic speech and is an essential factor for natural-sounding prosody.

Different approaches have been proposed for intensity modeling, ranging from rule-based approaches to statistical methods and deep learning techniques. Rule-based methods rely on linguistic and phonetic rules to generate intensity patterns, while statistical methods, such as Gaussian mixture models or hidden Markov models, learn the relationship between linguistic features and intensity from data. Recently, deep learning methods, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been employed for intensity modeling, leveraging their capacity to learn complex patterns in the data [80].

D. Pause Modeling

Pauses play a crucial role in speech synthesis, as they help convey the structure of spoken language, provide time for the listener to process information, and contribute to the naturalness of synthetic speech. Pause modeling involves predicting the timing and duration of pauses in speech synthesis.

Many techniques have been proposed for pause modeling, including rule-based approaches, statistical methods, and deep learning techniques. Rule-based methods rely on linguistic and syntactic rules to predict pause locations and durations, while statistical methods learn these relationships from data [89]. Deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), can also be employed for pause modeling, as they are capable of learning complex patterns and capturing long-range dependencies in the data.

VII. CHILDREN SPEECH SYNTHESIS

Children's speech synthesis is the process of creating artificial voices that sound like children, which is useful in developing interactive systems and robots for children's education and entertainment [81], [90]. However, this area of research poses several challenges. First, obtaining high-quality and phonetically balanced speech data from children is difficult. Additionally, children's voices have distinct characteristics that set it apart from adult speech. Mispronounced words, disfluencies, and ungrammatical utterances often characterize child speech. Furthermore, children exhibit linguistic differences compared to adult speech across different levels, such as prosody, vocabulary, grammar, and sizeable acoustic variability of child speech [91]. Moreover, synthesizing expressive conversational speech is a further challenge, as it requires the inclusion of paralinguistics and emotions in the synthesized speech [92]. Evaluating the quality of children's speech synthesis is also not straightforward, as it involves prolonged exposure to the synthetic voice.

Despite these challenges, researchers are exploring various approaches, such as speaker adaptive. a study conducted by [93] explored the acoustic characteristics of children's speech, encompassing aspects such as duration and pitch the results indicated that certain vowel sounds have longer durations in children compared to adults. Moreover, synthesizing expressive conversational speech is a further challenge, as it requires the inclusion of paralinguistics and emotions in the synthesized speech [92].

Evaluating the quality of children's speech synthesis is also not straightforward, as it involves prolonged exposure to the synthetic voice. Despite these challenges, researchers are exploring various approaches, such as speaker-adaptive HMM-based speech synthesis and deep learning techniques, to develop efficient and accurate methods for children's speech synthesis.

The goal is to make dialogue systems more inclusive and accessible for younger users. Hidden Markov Models (HMMs) have been used in child speech synthesis to find suitable initial models and speaker adaptation methods [94], [95]. Nevertheless, HMM-based systems for synthesizing child speech often face difficulties in achieving high naturalness and accurately replicating the subtleties of children's speech. In this study [91],

The researchers introduced deep neural vocoders within a TTS framework to achieve child speech synthesis. Their method involves fine-tuning both the acoustic model Tacotron2 and a pre-trained WaveRNN vocoder. Moreover, they performed additional fine-tuning of the WaveRNN vocoder on a dedicated child speech dataset, improving the quality of child speech

synthesis [96]. In [97], a hybrid system that combines DNN with HMM was utilized for automatic speech recognition, using approximately 10 hours of Italian child speech data. This hybrid DNN-HMM approach proved effective in enhancing speech recognition accuracy specifically for Italian child speech.

VIII. Discussion and Conclusion

Speech synthesis has come a long way since the early days of simple rule-based systems. Today, there are a variety of approaches and techniques that can be used to generate natural-sounding synthetic speech. This survey offers an overview of the development of expressive Text-to-Speech (TTS) systems and the diverse methodologies employed to synthesize expressive speech from written text. The selected articles presented a range of TTS and speech synthesis models that aim to enhance the quality and expressiveness of synthetic speech. This survey encapsulates the contemporary as well as conventional methods that are utilized in TTS systems. We discussed deep learning-based speech synthesis, emotional speech synthesis and style transfer in speech synthesis. Additionally, we have reviewed several objective metrics such as Itakura-Saito measure, Root mean square (RMSE), Gross pitch error (GPE) and subjective metrics such as MOS and MUSHRA utilized to access the quality of the synthesized speech are examined. In addition, our focus was on the representation and transfer approaches for style to comprehensively illustrate the significance of style representation in enhancing the expressiveness of synthesized speech in Text-to-Speech (TTS) systems. Further, we reviewed both deep learning-based autoregressive model such as Parallel WaveNet and non-autoregressive model such as FastSpeech that are used in the front-end and back-end of TTS system.

Finally, we point out the challenges in child speech synthesis, which involves the difficulty of obtaining high-quality and phonetically balanced speech data from children. Additionally, we address the unique characteristics of children's speech, differentiating it from adult speech, including linguistic variations and expressive conversational patterns.

We hope this paper will offer a clear overview for readers to understand the current status of expressive speech synthesis models, inspiring continuous research efforts on expressive TTS systems. This, in turn, aims to promote future modern in the field of study expressive TTS systems, especially in the field of child speech synthesis.

IX. Acknowledgements

This paper is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and by the Ministry of Innovation and Culture and the National Research, Development and Innovation Office of Hungary within the framework of the National Laboratory of Artificial Intelligence. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union and the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.

References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A Survey on Neural Speech Synthesis," Jun. 2021, [Online]. arXiv preprint *arXiv:2106.15561*, 2021. Available: http://arxiv.org/abs/2106.15561.

[2] N. Tits, "Controlling the emotional expressiveness of synthetic speech: a deep learning approach," *4OR*, vol. 20, no. 1, pp. 165–166, Mar. 2022, DOI: 10.1007/s10288-021-00473-2.

[3] P. Alexander. Taylor, Text-to-speech synthesis. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511816338

[4] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, "Review of deep learning-based speech synthesis," *Applied Sciences (Switzerland)*, vol. 9, no. 19. MDPI AG, Oct. 01, 2019. DOI: 10.3390/app9194050.

[5] K. R. Scherer, "Vocal affect expression: a review and a model for future research.," *Psychol Bull*, vol. 99, no. 2, p. 143, 1986. DOI: 10.1037/0033-2909.99.2.143

[6] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for american english," *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 4, pp. 1099–1108, Jul. 2006, DOI: 10.1109/TASL.2006.876123.

[7] M. Mahrishi, K. K. Hiran, G. Meena, and P. Sharma, Machine learning and deep learning in real-time applications. IGI global, 2020. DOI: 10.4018/978-1-7998-3095-5.ch009

[8] D. H. Klatt, "Review of text-to-speech conversion for English," *J Acoust Soc Am*, vol. 82, no. 3, pp. 737–793, 1987. DOI: 10.1121/1.395275

[9] D. H. Klatt, "Software for a cascade/parallel formant synthesizer." *Acoust Soc Am*, vol. 67, no. 3, pp. 971–995, 1980, DOI: 10.1121/1.383940

[10] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 4, pp. 1145–1153, Jul. 2006, DOI: 10.1109/tasl.2006.876113

[11] N. Campbell, W. Hamza, H. Hoge, J. Tao, and G. Bailly, "Editorial Special Section on Expressive Speech Synthesis," *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 4, pp. 1097–1098, Jun. 2006, DOI: 10.1109/tasl.2006.878306.

[12] D. Govind and S. R. M. Prasanna, "Expressive speech synthesis: a review," *Int J Speech Technol*, vol. 16, pp. 237–260, 2013. DOI: 10.1007/s10772-012-9180-2

[13] C. Valentini-Botinhao, M. S. Ribeiro, O. Watts, K. Richmond, and G. E. Henter, "Predicting pairwise preferences between TTS audio stimuli using parallel ratings data and anti-symmetric twin neural networks," *International Speech Communication Association, INTERSPEECH*, 2022. DOI: 10.21437/interspeech.2022-10132

[14] International Telecommunication Union (ITU), "Methods for subjective determination of transmission quality," 1996. DOI: 10.18356/16e04175-en.

[15] H. Zen, A. Senior, and M. S. Google, "Statistical Parametric Speech Synthesis Using Deep Neural Networks." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2013. DOI: 10.1109/icassp.2013.6639215

[16] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," Sep. 2016, arXiv preprint *arXiv:1609.03499*, DOI: 10.48550/arXiv.1609.03499

[17] Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," *International Speech Communication Association, INTERSPEECH*, 2017, DOI: 10.21437/interspeech.2017-1452

[18] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191.

[19] X. Zhuang, T. Jiang, S. Y. Chou, B. Wu, P. Hu, and S. Lui, "Litesing: Towards Fast, Lightweight And Expressive Singing Voice Synthesis," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7078–7082. DOI: 10.1109/icassp39728.2021.9414043.

[20] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An Investigation Of Subband Wavenet Vocoder Covering Entire Audible Frequency Range With Limited Acoustic Features," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5654–5658. DOI: 10.1109/icassp.2018.8462237

[21] A. Van Den Oord et al., "Parallel WaveNet: Fast High-Fidelity Speech Synthesis," In *International conference on machine learning* (pp. 3918–3926). PMLR, 2018. arXiv preprint *arXiv:1711.10433* DOI: 10.48550/arXiv.1711.10433

[22] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal Neural Vocoding with Parallel Wavenet," In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6044–6048. 6048. DOI: 10.1109/icassp39728.2021.9414444

[23] H. Choi, S. Park, J. Park, and M. Hahn, "Emotional Speech Synthesis For Multi-Speaker Emotional Dataset Using Wavenet Vocoder," In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2019, pp. 1–2. DOI: 10.1109/icce.2019.8661919

[24] J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, "Autovocoder: Fast Waveform Generation from a Learned Speech Representation Using Differentiable Digital Signal Processing," In *ICASSP – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095729.

[25] D. Griffin and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans Acoust*, vol. 32, no. 2, pp. 236–243, 1984. DOI: 10.1109/tassp.1984.1164317

[26] J. Kong, J. Kim, and J. Bae, "Hifi-Gan: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Adv Neural Inf Process Syst*, vol. 33, pp. 17022–17033, 2020. https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

[27] S. Li et al., "SnakeGAN: A Universal Vocoder Leveraging DDSP Prior Knowledge and Periodic Inductive Bias," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 1703–1708. DOI: 10.1109/icme55011.2023.00293

[28] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP 2020, - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 6199–6203. DOI: 10.1109/ICASSP40776.2020.9053795.

[29] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A Universal Neural V ocoder With Large-Scale Training,", 2022. DOI: 10.48550/arXiv.2206.04658

[30] Y. Zheng, J. Tao, Z. Wen, and J. Yi, "Forward-backward decoding sequence for regularizing end-to-end TTS," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 27, no. 12, pp. 2067–2079, Dec. 2019, DOI: 10.1109/taslp.2019.2935807.

[31] J. Shen et al., "Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783. https://doi.org/10.1109/icassp.2018.8461368

[32] A. Kulkarni, V. Colotte, and D. Jouvet, "Improving Transfer of Expressivity For End-To-End Multispeaker Text-To-Speech Synthesis." In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 31–35. IEEE, 2021. DOI: 10.23919/eusipco54536.2021.9616249

[33] R. J. Skerry-Ryan et al., "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *ICML* 2018. http://proceedings.mlr.press/v80/skerry-ryan18a.html

[34] Y. Ren et al., "Fastspeech: Fast, Robust And Controllable Text To Speech," *Advances in neural information processing systems Adv Neural Inf Process Syst*, vol. 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html

[35] A. Łańcucki, "Fastpitch: Parallel Text-To-Speech With Pitch Prediction," in I*CASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592. DOI: 10.1109/icassp39728.2021.9413889

[36] Y. Ren et al., "Fastspeech 2: Fast And High-Quality End-To-End Text To Speech," 2020, arXiv preprint *arXiv:2006.04558*, DOI: 10.48550/arXiv.2006.04558.

[37] P . C. Loizou, "Speech quality assessment." *Multimedia analysis, processing and communications 2011*, pp. 623–654. DOI: 10.1007/978-3-642-19551-8_23

[38] B. Sabine, J. Latorre, and K. Yanagisawa. "Crowdsourced assessment of speech synthesis." Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment (2013): 173–216, DOI: 10.1002/9781118541241.ch7

[39] I. Recommendation, "1534-1,'Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA),'" *International Telecommunications Union*, Geneva, Switzerland, vol. 2, 2001. https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf

[40] B.-H Juang, "On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 8, pp. 1477–1498, 1984, DOI: 10.1002/j.1538-7305.1984.tb00047.x

[41] C.-C. Wang, Z.-H. Ling, B.-F. Zhang, and L.-R. Dai, "Multi-Layer F0 Modeling For HMM-Based Speech Synthesis," in *2008 6th International symposium on Chinese spoken language processing*, IEEE, 2008, pp. 1–4. DOI: 10.1109/chinsl.2008.ecp.44

[42] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A Comparative Study of Pitch Extraction Algorithms on a Large Variety of Singing Sounds," Dec. 2019, DOI: 10.1109/icassp.2013.6639185

[43] Y. Wang et al., "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *ICML* 2018. https://proceedings.mlr.press/v80/wang18h.html?ref=https://githubhelp.com

[44] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting Expressive Speaking Style From Text In End-To-End Speech Synthesis," In *IEEE Spoken Language Technology Workshop (SLT)*, 2018. DOI: 10.1109/slt.2018.8639682

[45] Y. Wang et al., "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," 2018. https://proceedings.mlr.press/v80/wang18h.html

[46] S. Liu, S. Yang, D. Su, and D. Yu, "Referee: Towards Reference-Free Cross-Speaker Style Transfer With Low-Quality Data For Expressive Speech Synthesis," In *ICASSP ,IEEE International Conference on Acoustics Speech and Signal , Processing (ICASSP)*, pp. 6307–6311. IEEE, 2022. DOI: 10.1109/icassp43922.2022.9746858

[47] C. Yu et al., "DurIAN: Duration Informed Attention Network for Multimodal Synthesis," *International Speech Communication Association, INTERSPEECH*, 2019, pp. 2027–2031, DOI: 10.21437/interspeech.2020-2968

[48] Y. Lee and T. Kim, "Robust And Fine-Grained Prosody Control Of End-To-End Speech Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5911–5915. IEEE, 2019. DOI: 10.1109/icassp.2019.8683501

[49] W. Lu et al., "One-Shot Emotional Voice Conversion Based On Feature Separation," *Speech Commun*, vol. 143, pp. 1–9, Sep. 2022, DOI: 10.1016/j.specom.2022.07.001.

[50] D. Wang, L. Li, Y. Shi, Y. Chen, and Z. Tang, "Deep Factorization for Speech Signal," " *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5094–5098. IEEE, 2018. DOI: 10.1109/icassp.2018.8462169

[51] J. Wang, J. Li, X. Zhao, Z. Wu, S. Kang, and H. Meng, "Adversarially Learning Disentangled Speech Representations For Robust Multi-Factor Voice Conversion," *International Speech Communication Association, INTERSPEECH*, 2021, pp. 846–850, DOI: 10.21437/interspeech.2021-1990

[52] J. Chou, C. Yeh, H. Lee, and L. Lee, "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations," *International Speech Communication Association, INTERSPEECH*, 2018, pp. 501–505, DOI: 10.21437/interspeech.2018-1830

[53] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving Zero-shot V oice Style Transfer via Disentangled Representation Learning," *International Conference on learning representation* 2021, https://openreview.net/forum?id=TgSVWXw22FQ

[54] Y. Shin, Y. Lee, S. Jo, Y. Hwang, and T. Kim, "Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 2313–2317, DOI: 10.21437/interspeech.2022-10131

[55] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference Tacotron by Intercross Training for Style Disentangling,Transfer and Control in Speech Synthesis," Apr. 2019, arXiv preprint *arXiv:1904.02373* DOI: 10.48550/arXiv.1904.02373

[56] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-Reference Neural TTS Stylization with Adversarial Cycle Consistency," *International Speech Communication Association, INTERSPEECH*, 2020, pp. 4442–4446, DOI: 10.21437/interspeech.2020-2985

[57] S. Pan, "Cross-speaker Style Transfer with Prosody Bottleneck in Neural Speech Synthesis." *International Speech Communication Association, INTERSPEECH*, 2021, pp. 4678–4682 DOI: 10.21437/interspeech.2021-979

[58] X. Li, C. Song, X. Wei, Z. Wu, J. Jia, and H. Meng, "Towards Cross-speaker Reading Style Transfer on Audiobook Dataset," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 5528–5532, DOI: 10.21437/interspeech.2022-11223

[59] J. Zaïdi, H. Seuté, B. van Niekerk, and M.-A. Carbonneau, "Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 4591–4595, DOI: 10.21437/interspeech.2022-10761

[60] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual Reasoning with A General Conditioning Layer," in *Proceedings of the AAAI conference on artificial intelligence*, 2018. DOI: 10.1609/aaai.v32i1.11671

[61] R. Terashima et al., "Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 3018–3022, DOI: 10.21437/interspeech.2022-11278

[62] Z. Shang, Z. Huang, H. Zhang, P. Zhang, and Y. Yan, "Incorporating Cross-Speaker Style Transfer For Multi-Language Text-To-Speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 3406–3410. DOI: 10.21437/Interspeech.2021-1265.

[63] P. Wu et al., "Cross-speaker Emotion Transfer Based on Speaker Condition Layer Normalization and Semi-Supervised Training in Text-To-Speech," Oct. 2021, arXiv preprint *arXiv:2110.04153*, DOI: 10.48550/arXiv.2110.04153

[64] C. Qiang, P. Yang, H. Che, X. Wang, and Z. Wang, "Style-Label-Free: Cross-Speaker Style Transfer by Quantized VAE and Speaker-wise Normalization in Speech Synthesis," Dec. 2022, in *13th International Symposium on Chinese Spoken Language Processing (ISCSLP)* Dec. 2022, DOI: 10.1109/iscslp57327.2022.10038135

[65] S. Aryal and R. Gutierrez-Osuna, "Accent Conversion Through Cross-Speaker Articulatory Synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 7694–7698. DOI: 10.1109/icassp.2014.6855097

[66] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-Resource Expressive Text-To-Speech Using Data Augmentation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6593–6597. IEEE, 2021. DOI: 10.1109/icassp39728.2021.9413466

[67] J. Wu, A. Polyak, Y. Taigman, J. Fong, P. Agrawal, and Q. He, "Multilingual Text-To-Speech Training Using Cross Language Voice Conversion And Self-Supervised Learning Of Speech Representations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8017–8021. DOI: 10.1109/icassp43922.2022.9746282

[68] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Multi-Speaker TTS System For Low-Resource Language Using Cross- Lingual Transfer Learning And Data Augmentation," in *2021 Asia- Pacific Signal and Information Processing Association Annual Summit and Conference (APSIP A ASC)*, IEEE, 2021, pp. 849–853. https://ieeexplore.ieee.org/abstract/document/9689505

[69] Z. Zhang, Y. Zheng, X. Li, and L. Lu, "WeSinger: Data-augmented Singing Voice Synthesis with Auxiliary Losses," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 4252–4256, DOI: 10.21437/interspeech.2022-454

[70] Y. Nakai, Y. Saito, K. Udagawa, and H. Saruwatari, "Multi-Task Adversarial Training Algorithm for Multi-Speaker Neural Text-to- Speech," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 743–748 DOI: 10.23919/apsipaasc55919.2022.9980331

[71] X. Zhang, J. Wang, N. Cheng, and J. Xiao, "TDASS: Target Domain Adaptation Speech Synthesis Framework for Multi-speaker Low-Resource TTS," International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE, 2022, DOI: 10.1109/ijcnn55064.2022.9892596

[72] K. Inoue, S. Hara, and M. Abe, "Module Comparison of Transformer-TTS For Speaker Adaptation Based On Fine-Tuning," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIP A ASC)*, IEEE, 2020, pp. 826–830. https://ieeexplore.ieee.org/abstract/document/9306250

[73] C. Du, Y. Guo, X. Chen, and K. Yu, "Speaker Adaptive Text-to-Speech with Timbre-Normalized Vector-Quantized Feature," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023). DOI: 10.1109/taslp.2023.3308374

[74] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Speaker Adaptation Experiments with Limited Data for End-to-End Text-To-Speech Synthesis using Tacotron2," *Infocommunications Journal*, vol. 14, no. 3, pp. 55–62, 2022. DOI: 10.36244/icj.2022.3.7

[75] C.-P. Hsieh, S. Ghosh, and B. Ginsburg, "Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers," *International Speech Communication Association, INTERSPEECH*, 2023, pp. 3028–3032 DOI: 10.21437/interspeech.2023-2313

[76] Y. Jia et al., "Transfer Learning from Speaker V erification to Multi-speaker Text-To-Speech Synthesis." *Advances in neural information processing systems*, 2018. https://proceedings.neurips.cc/paper_files/paper/2018/hash/6832a7b24bc06775d02b7406880b93fc-Abstract.html

[77] K. Inoue, S. Hara, M. Abe, T. Hayashi, R. Yamamoto, and S. Watanabe, "Semi-Supervised Speaker Adaptation For End-To-End Speech Synthesis With Pretrained Models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7634–7638. IEEE, 2020. DOI: 10.1109/icassp40776.2020.9053371

[78] M. Zhang, X. Zhou, Z. Wu, and H. Li, "Towards Zero-Shot Multi-Speaker Multi-Accent Text-to-Speech Synthesis," *IEEE Signal Processing Letters* 2023. DOI: 10.1109/lsp.2023.3292740

[79] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, "BOFFIN TTS: Few-Shot Speaker Adaptation by Bayesian Optimization," In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7639–7643. IEEE, 2020. DOI: 10.1109/icassp40776.2020.9054301

[80] J. P. H. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, Progress in speech synthesis. Springer Science & Business Media, 2013. DOI: 10.1007/978-1-4612-1894-4_15

[81] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," *Artif Intell Rev*, vol. 56, no. 7, pp. 5837–5880, Jul. 2023, DOI: 10.1007/s10462-022-10315-0.

[82] Y. J. Zhang, C. Zhang, W. Song, Z. Zhang, Y. Wu, and X. He, "Prosody Modelling with Pre-Trained Cross-Utterance Representations for Improved Speech Synthesis," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 31, pp. 2812–2823, 2023, DOI: 10.1109/TASLP.2023.3278184.

[83] H. Bae and Y.-S. Joo, "Enhancement of Pitch Controllability using Timbre-Preserving Pitch Augmentation in FastPitch," *International Speech Communication Association, INTERSPEECH*, 2022, pp. 6–10 DOI: 10.21437/interspeech.2022-55

[84] N. Adiga and S. R. M. Prasanna, "Acoustic Features Modelling for Statistical Parametric Speech Synthesis: A Review," *IETE Technical Review*, vol. 36, no. 2, pp. 130–149, 2019. DOI: 10.1080/02564602.2018.1432422

[85] J. Ni, Y. Shiga, and H. Kawai, "Duration Modeling with Global Phoneme- Duration Vectors.," I*nternational Speech Communication Association, INTERSPEECH*, 2019, pp. 4465–4469. DOI: 10.21437/interspeech.2019-2126

[86] J. Lian, C. Zhang, G. K. Anumanchipalli, and D. Yu, "Unsupervised TTS Acoustic Modeling for TTS with Conditional Disentangled Sequential V AE," *IEEE/ACM Trans Audio Speech Lang Process*, 2023. DOI: 10.1109/taslp.2023.3290423

[87] K. Fujita, T. Ashihara, H. Kanagawa, T. Moriya, and Y. Ijima, "Zero-Shot Text-To-Speech Synthesis Conditioned Using Self-Supervised Speech Representation Model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* 2023. DOI: 10.1109/icasspw59220.2023.10193459

[88] A. Abbas et al., "Expressive, variable, and controllable duration modelling in TTS," *International Speech Communication Association, INTERSPEECH*, 2022. DOI: 10.21437/interspeech.2022-384

[89] Y. Stylianou, "Applying the Harmonic Plus Noise Model in *Concatenative Speech Synthesis*," 2001. DOI: 10.1109/89.890068

[90] J. Y. Zhang, A. W. Black, and R. Sproat, "Identifying Speakers in Children's Stories for Speech Synthesis." In *Eighth European Conference on Speech Communication and Technology*. 2003. DOI: 10.21437/eurospeech.2003-586

[91] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," *IEEE Access*, vol. 10, pp. 47 628–47 642, 2022, DOI: 10.1109/access.2022.3170836

[92] A. Borgh, K. ; Dickson, W. Patrick, K. Borgh, and W. P. Dickson, "DOCUMENT RESUME ED 277 007 CS 210 188 The Effects on Children's Writing of Adding Speech Synthesis "permission to reproduce this material has been granted by," 1986. DOI: 10.1080/08886504.1992.10782629

[93] C. Terblanche, M. Harty, M. Pascoe, and B. V Tucker, "A Situational Analysis of Current Speech-Synthesis Systems for Child Voices: A Scoping Review of Qualitative and Quantitative Evidence," *Applied Sciences*, vol. 12, no. 11, p. 5623, 2022.

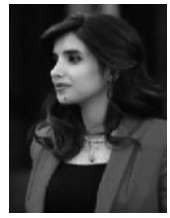
[94] A. Govender, F. de Wet, and J.-R. Tapamo, "HMM Adaptation For Child Speech Synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. DOI: 10.21437/interspeech.2015-379

[95] A. Govender and F. De Wet, "Objective Measures To Improve The Selection Of Training Speakers In HMM-Based Child Speech Synthesis," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, IEEE, 2016, pp. 1–6. DOI: 10.1109/robomech.2016.7813193

[96] D. Giuliani and B. BabaAli, "Large Vocabulary Children's Speech Recognition with DNN-HMM and SGMM Acoustic Modeling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. DOI: 10.21437/interspeech.2015-378

[97] P. Cosi, "A Kaldi-Dnn-Based Asr System For Italian," in *International Joint Conference On Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–5. DOI: 10.1109/ijcnn.2015.7280336

[98] Matsumoto, Kento, Sunao Hara, and Masanobu Abe. "Speech-Like Emotional Sound Generation Using WaveNet." *IEICE TRANSACTIONS on Information and Systems* 105, no. 9 (2022): 1581–1589. DOI: 10.1587/transinf.2021edp7236



Shaimaa Alwaisi was born in, Iraq. She got a BSc degree in Computer Engineering at Diyala University, higher Diploma from Iraqi commission for computer and informatics ICCI and a MSc degree in Computer Engineering at Selcuk University, Turkey. She currently PhD student at the Speech Technology and Smart Interactions Laboratory in the Budapest University of Technology and Economics. She is working on neural vocoders and acoustic models for speech synthesis. her current interests are signal processing, expressive speech synthesis, Child speech synthesis, Deep learning, acoustic models, and voice conversion.

Géza Németh was born in 1959. He obtained his MSc in electrical engineering, major in Telecommunications at the Faculty of Electrical Engineering of BME in 1983. Also, at BME: dr. univ., 1987, PhD 1997. He is an associate professor at BME. He is the author or co-author of more than 170 scientific publications and 4 patents. His research fields include speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications. He is the Head of the Speech Technology and Smart Interactions Laboratory of BME TMIT.