An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era

By Andreas Triantafyllopoulos, Björn W. Schuller, *Fellow IEEE*, Gökçe İymen, Metin Sezgin, *Member IEEE*, Xiangheng He, Zijiang Yang, *Student Member IEEE*, Panagiotis Tzirakis, *Member IEEE*, Shuo Liu, Silvan Mertes, Elisabeth André, *Senior Member IEEE*, Ruibo Fu, *Member IEEE*, and Jianhua Tao, *Senior Member IEEE*

ABSTRACT | Speech is the fundamental mode of human communication, and its synthesis has long been a core priority in human–computer interaction research. In recent years, machines have managed to master the art of generating speech that is understandable by humans. However, the linguistic content of an utterance encompasses only a part of its meaning. Affect, or expressivity, has the capacity to turn speech into a medium capable of conveying intimate thoughts, feelings, and emotions—aspects that are essential for engaging and naturalistic interpersonal communication. While the goal of imparting expressivity to synthesized utterances has so far remained elusive, following recent advances in text-to-speech synthesis, a paradigm shift is well under way in the fields of affective speech synthesis and conversion as well. Deep learning, as the technology that underlies most of the recent advances in artificial intelligence, is spearheading these efforts. In this overview, we outline ongoing trends and summarize state-of-the-art approaches in an attempt to provide a broad overview of this exciting field.

KEYWORDS | Affective computing; deep learning; emotional voice conversion (EVC); speech synthesis.

Andreas Triantafyllopoulos, **Xiangheng He**, **Zijiang Yang**, and **Shuo Liu** are with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: andreas.triantafyllopoulos@uni-a.de).
Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the Group on Language, Audio, & Music (GLAM), Imperial College London, SW7 2AZ London, U.K.
Gökçe İymen and **Metin Sezgin** are with the KUIS AI Laboratory, College of Engineering, Koç University, 34450 Istanbul, Turkey.
Panagiotis Tzirakis is with the Group on Language, Audio, & Music (GLAM), Imperial College London, SW7 2AZ London, U.K.
Silvan Mertes and **Elisabeth André** are with the Chair of Human-Centered Artificial Intelligence, University of Augsburg, 86159 Augsburg, Germany.
Ruibo Fu and **Jianhua Tao** are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

I. INTRODUCTION

> We all have the capacity to be creative. We're all driven to share our deepest dreams and ideas with the world. When we think of the most talented creative people, they speak to us in a unique way. A phrase we often hear is *"Having a creative voice."*
>
> — Val Kilmer

The story of Val Kilmer, a world-renowned actor who lost his voice to throat cancer at the peak of his career, is a poignant reminder of the importance of verbal communication in human societies.[1] A voice is more than the sum of its words; it is a conduit of one's individuality,

[1] A video of the reconstruction of Val Kilmer's voice for the purposes of Top Gun 2 by SONANTIC can be found in https://www.youtube.com/watch?v=OSMue60Gg6s

emotions, and unique worldview. People who suffer from similar conditions understand that the mere verbalization of their words using assistive technologies is not enough to give them back their voices. They need to regain their lost *emotional expressivity* [1].

If artificial beings are ever able to attain an equal standing in human societies, why should they need any less? While contemporary artificial intelligence (AI) research has set its sights on more attainable, down-to-earth goals, the long-standing dream of AI researchers is to simulate, or perhaps overcome, human intelligence. This goal may well require machines to have emotions as, to quote one of the forefathers of the field, Minsky [2]: "the question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions." Any entity that has emotions requires an avenue to express them.

Affective computing is the subfield of AI that concerns itself with the computational modeling, understanding, and expression of emotions [3]. One of its primary goals is to facilitate more natural human–computer interaction (HCI) through the modeling of affect, which is a key component of human behavior. To that end, language, and, in particular, *spoken* language, is the most natural form of communication. If machines are ever to become natural conversational partners, they have to master the art of speech generation—including the prosodic intonations attributable to the expression of affect [4]. This is the domain of *affective speech synthesis*, a computational paradigm that attempts to generate realistic-sounding affective speech. We define affective speech synthesis as a subfield of voice transformation [5], which corresponds to the modification of all potential parameters of speech, and as a super-field of emotional speech synthesis (ESS), which corresponds to the modification of emotion. Affective speech synthesis, in contrast, extends beyond emotions by covering all aspects that fall under the umbrella of *computational paralinguistics* [6], such as mood, personality, and social status. Nevertheless, as emotion is heavily overrepresented in recent deep learning (DL)-based affective speech synthesis works, our review will be largely geared toward that particular type of synthesis. We also consider both the general case of synthesizing an affective utterance directly from the input text, as well as that of modifying a neutral one to capture the desired emotion—a subfield of synthesis generally referred to as affective, or emotional, voice conversion.

The first attempts to infuse emotion into synthesized speech were made before the field's name was even coined [7], [8], [9], [10], [11], [12]. For a long period, research on emotional speech synthesis and conversion has primarily focused on rule-based approaches guided by experts and listening experiments.[2] This is in contrast to speech emotion recognition (SER)—the 'opposite end' of synthesis—which has been dominated by a data-driven paradigm [13].

The last few years have seen tremendous progress in the "sister fields" of speech synthesis and voice conversion. The landmark work of van den Oord et al. [14] revolutionized the field of text-to-speech synthesis (TTS), signaling the advent of the DL era, and, more generally, solidifying the switch to a data-driven paradigm, where a mapping from text to speech is *learned* using data. Similar approaches are now spearheading research in affective speech synthesis as well [13].

The TTS approaches have reached such performance levels that the task is considered by many to be "solved"—layman users in particular expect commercial TTS systems to work flawlessly, as seen, for example, in the recent wave of voice assistants. Accordingly, TTS research has exploded in recent years. In contrast, the field of affective speech synthesis has attracted somewhat less attention in the field of HCI, which is, nevertheless, substantially increasing. Yet, even though significant progress has been made in that area as well, the goal of human-level, controllable emotional expressivity still remains elusive.

In an attempt to summarize recent efforts, synthesize existing approaches, identify missing gaps, and highlight promising research directions, we have construed a literature review of *deep*, *affective* speech synthesis, and conversion methods. Our overview, thus, fills the gap between recent surveys in deep speech synthesis, which focuses on "mere" TTS [15], and older affective speech synthesis reviews that have become largely obsolete in the deep learning era [12], [16] or newer ones that are more limited in scope [17], [18].

Our review aims to provide a broad overview of different aspects of affective speech synthesis. This is informed by the authors' own knowledge and understanding of the field (and, thus, unavoidably, our biases). However, to provide a more balanced overview of the different *technical* aspects of DL-based methods, we also conducted a more systematic review of existing literature. We limited our search to articles published after January 1, 2012, and until August 31, 2022. We identified a total of 101 relevant works using IEEE Xplore (76) and Web of Science (25) and searching for articles with one of the following terms in their title: "affective speech synthesis" (4), "emotional speech synthesis" (37), "emotional voice conversion (EVC)" (19), "expressive voice conversion" (1), "expressive text-to-speech" (15), and "expressive speech synthesis" (27). From these, we excluded seven for referring to synthesis for multimodal data (gestures/video), one for being written in a non-English language, two for being inaccessible to us, two for introducing data resources, and 41 for using non-deep neural network (DNN)-based systems. Furthermore, aside from [19], [20], and [21] which generate emphatic speech, all other works (45) focus on generating emotional speech. This justifies our choice to make this one aspect of affect a major focal point of our review, especially with regards to Section IV, which outlines the technical con-

[2] An online "museum" including listening examples of most such attempts is found at http://emosamples.syntheticspeech.de/

siderations behind DNN architectures. With that in mind, the identified works were complemented by searching in related references and adding works familiar to the authors but not found with the above search criteria.

The remainder of this work is structured as follows. We first present an overview of where affective speech synthesis fits in an affective computing application in Section II. We then give a brief introduction on (deep) speech synthesis in general in Section III, followed by a thorough, technical review of deep emotional speech synthesis, and conversion (see Section IV). Finally, we summarize major observations and outline potential avenues for future research in Section V.

II. AFFECTIVE SPEECH GENERATION

In Section II, we begin by giving a definition of what is entailed by affective speech generation. We use the term "generation" to encompass all aspects of a process that begins with a "decision" on what *emotion, style, or stance* needs to be generated, a selection of the appropriate text, and the final *synthesis* of the waveform as the last step. As affect is an overloaded term, we first give a concrete definition of it for the purposes of our review. After defining what we mean by it, we continue our overview with a short introduction to the implicit model, which underlies all affective computing applications: that of an agent who is able to respond emotionally to external stimuli. This agent may be fully artificial, as in the case of an autonomous conversational agent that interacts with humans, or "hybrid" in the case of an EVC system that augments the capabilities of speech-impaired individuals [1]. We then introduce the module of that agent responsible for the generation of affect in speech, followed by an introduction of the different representation models used in the computational modeling of affect. Finally, we introduce the acoustic correlates of affect, which have guided related speech research for decades.

A. What Is Affective Speech?

We begin with a definition of the concepts used here as, due to their subjective nature, they are often conflated with one another. We use the term "affect" in its broadest connotation, as it was introduced in the inaugural work of Picard [3]. Specifically, we go by the definition of affect "as a broader term, encompassing all kinds of manifestations of personality such as mood, interpersonal stances, or attitudes" [6]. This differs from standard psychological interpretations that more narrowly define affect as the manifestation of a subjectively experienced emotion [22]. Thus, in our review, affective speech is speech that encapsulates all possible paralinguistic traits and states [6].

Despite the importance of those other aspects of affect, the majority of recent research in affective speech synthesis has been actually devoted to ESS, with considerably less emphasis on personality and other states or traits [6], [23]. This discrepancy is even more pronounced in the ongoing deep learning era, with far more work devoted to ESS than any other construct.

We further note that several works are investigating explicit prosodic or rhythmic control [24], [25], [26], [27], [28]. Even though prosody and rhythm are critical components of emotional speech, they are not the only ones [29]. Moreover, it is often the case that this control is applied manually to change the speaking style of the synthesized utterance. It is, thus, missing the explicit link to affect that we consider critical for ESS. Naturally, prosody is highly related to it; therefore, any manipulation of prosody might result in a change in the perceived affect. Yet, oftentimes, these methods tend to leave out any evaluation of the affective content of the synthesized utterances and only focus on evaluating the controlled attributes. For this reason, we only tangentially refer to them in our review. We also note that, in the earliest affective speech synthesis papers, expressivity was considered a synonym to affect [9] or seen as a mechanism with which to express affect [30]. Therefore, we too use the terms "affective" and "expressive" speech synthesis interchangeably.

B. Affective Agent Model

While the majority of affective speech generation works are concerned with the task of endowing a synthesized utterance with appropriate emotional inflections, this is but the last step in the pipeline of an affective agent.[3] Fig. 1 shows a coarse model of what is at play in an affective computing application. The agent, rather than existing in a vacuum, is embedded in an environment (e.g., its application or, even, the entire world) and interacts with an interlocutor (usually a human; in the future, potentially other artificial entities) [23]. It receives inputs from this environment—including responses/queries by its interlocutor—and generates an appropriate response.

An important step in this process is the appraisal of all input stimuli. According to appraisal theory [31], [32], inputs from the environment are evaluated with respect to the agent's goals and concerns along several dimensions. For example, Ellsworth and Scherer [32] proposed novelty (how much new information was contained in the stimulus), intrinsic pleasantness/valence (how positive or negative the stimulus "feels" for the agent), relevance (how pertinent is the information to its goals), urgency (how fast it needs to respond), and power/control (how much is the situation under its control). While this list is not exhaustive, and alternative appraisal theories have been proposed over the years, we believe that it captures a core component of an affective agent, namely, that the appropriate text and affect to be synthesized have to be somehow defined. This includes accounting for the relationship between the agent and its interlocutor(s) and

[3] With the word "agent" here, we mean a software component that simulates a desired behavior in any digital application, not necessarily an embodied conversational agent.

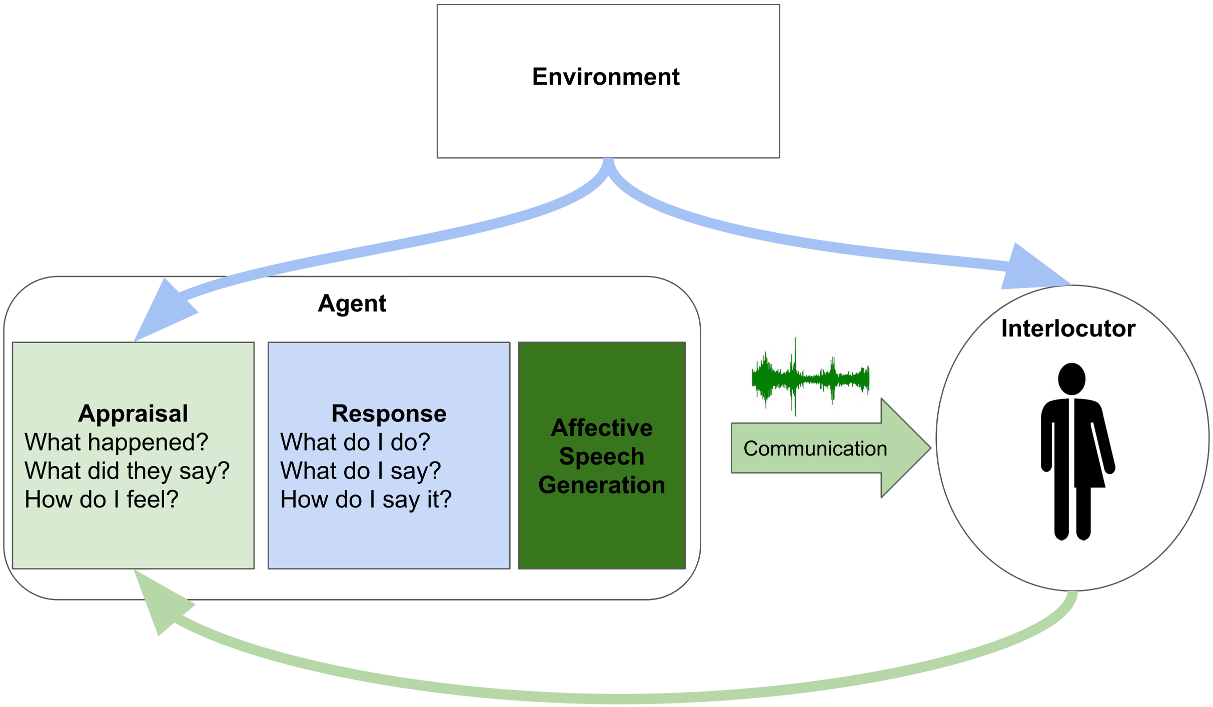


Fig. 1. ***Overview of our affective speech generation model. We assume the presence of an artificial agent who can receive inputs from the environment (including responses from their interlocutor) and proceed to appraise the situation (an appraisal can be either hardcoded or learned) and generate an appropriate response. This response is then converted to an appropriate speech signal by an affective speech generation module and transmitted to the interlocutor.***

the role that the agent is required to play in a particular application.

Of course, in most contemporary affective computing applications, the appropriate response is dictated by the creator of the application. Most of them contain hardcoded behaviors that the agents must follow (e.g., be constantly "happy" or "pleasing"). However, some recent works are already experimenting with learned behaviors—for example, using reinforcement learning to train a dialog agent to incorporate emotional responses [33], as these have been shown to increase subjective scores of dialog richness [34], [35], [36]. As the field progresses, we expect more research toward less hardcoded and more learned (or emerging) behaviors.

C. Computational Models of Affect

As Fehr and Russell [37] famously wrote: "Everyone knows what an emotion is, until asked to give a definition. Then, it seems, no one knows," however, to generate an emotion, one must, nevertheless, have a proper representation of it. Several different emotion theories have emerged over the years, each focusing on different, but oftentimes related, aspects of emotion [38]. Two of those have dominated the computational modeling of emotion [13]: discrete emotion theories, where emotions are considered to fall under discrete categories, such as Ekman's big six [39], and dimensional ones, such as Russel's arousal, valence, and dominance [40]. Most affective speech synthesis works have adopted the first formulation and assume emotion to come in discrete categories, thus transforming one to the other (or neutral to one of them), while only a few pursue the synthesis of dimensional affect instead [41], [42], [43]. Our review will accordingly focus on categorical ESS methods, as these have dominated the ongoing DL era.

Similar to emotions, there are of course all the other states and traits that may be desirable for an affective agent [6]. For example, Scherer et al. [44] discussed mood, interpersonal stances, attitudes, and personality. Each of those must be separately operationalized, e.g., personality is typically evaluated using a five-factor model (OCEAN) [45]. It is only after the respective concept has been appropriately coded that it may be annotated and its synthesis subsequently learned by data.

D. Acoustic Correlates of Affect

Affective speech synthesis is concerned with adapting those constituents of a speech signal that conveys affective information. Thus, progress in this field depends on progress in the mirror field of affect recognition and analysis, where considerably more research has been invested in the last decades [13], [46], [47]. Speech parameters that are identified as being conducive to the recognition of affect in speech are readily co-opted by researchers to control affect during synthesis and vice versa [4].

A large body of literature has linked the manifestation of affect in a speech to suprasegmental features, such as prosody, voice quality (e.g., jitter and shimmer), spectral and energy features, and temporal patterns, such as tempo and pausing [46], [47], [48]. For example, anger was shown to correspond to a higher mean F0 and energy, while "hot" anger also induced a higher variability and range of F0 [46]. These features, in turn, became the main knobs twisted by early affective speech synthesis models to achieve their required results [7], [8], [9]. Recent approaches have attempted to substitute them with learned representations, in the hope that those are better able to capture emotional information [49], [50]. Nevertheless, those features have left their mark on affective speech synthesis research as several works still use them—in some form—to guide the generation of emotional utterances. These approaches, therefore, constitute the main focus of our review.

There has also been some interest in other aspects of vocalizations impacted by affect. For example, Tahon et al. [51] investigated the potential of generating emotional pronunciations to improve expressivity. More recently, Baird et al. [52] launched the Expressive Vocalisations Workshop and Competition (ExVo) to foster more research in the generation of realistic emotional "vocal bursts" [53]. Combining such approaches with the synthesis methods removed here has great potential to improve the expressivity and emotionality of generated utterances, and we will discuss the potential of such attempts in Section V-C.

III. SPEECH SYNTHESIS

The goal of a speech synthesis system, also known as text-to-speech synthesis, is to generate artificial, human-like speech from a given text input. Speech synthesis is, thus, naturally, the backbone of affective speech synthesis, as the generation of realistic-sounding utterances is a prerequisite for enhancing its expressivity. The first recorded TTS system is Wolfgang von Kempelen's 18th-century pipes and bellows machine, which was able to produce vowel and consonant approximations, which, when properly combined, allowed visitors in his lab to recognize certain words [54]. The field has obviously progressed a lot from those early origins with the introduction of digital technology. Earlier digital attempts at TTS include articulatory, formant, and concatenative synthesis. The field then inherited advances in statistical machine learning and transitioned to the statistical parametric speech synthesis (SPSS) paradigm, whose influence is still ripe throughout contemporary DL-based TTS systems.

As the development of ESS has developed in tandem with that of TTS, we considered a brief overview of early synthesis methods necessary. This is followed by a review of *deep* speech synthesis methods in Section III-B, which sets the tone for our deep affective speech synthesis overview presented in Section IV. Section III is concluded with an overview of (deep) voice conversion, an application field of voice transformation that has a lot in common with ESS [5].

A. Brief History of Speech Synthesis

The earliest (digital) TTS systems attempted to simulate the human articulatory system by creating models for the movement of lips, tongue, glottis, and vocal tract—thus not differing much in spirit from the mechanical apparatus of von Kempelen. This came to be known as *articulatory synthesis* [55]. This paradigm was met with severe challenges in the modeling of articulatory behavior and was abandoned for a simpler, source-filter model that lends itself better to parameter control: *formant synthesis* [56], [57]. This type of synthesis relies on a rule-based modification of the formant amplitudes and frequencies of an excitation signal to produce the required utterance. These rules are derived by linguistic analysis. While this system has more modularity than articulatory synthesis, the difficulty in identifying an appropriate set of rules has led to its abandonment in favor of data-driven paradigms.

To overcome the challenges associated with building a proper articulatory model or assembling a complete list of formant rules, the community next turned to *concatenative speech synthesis*, where the target utterance is constructed from a set of prerecorded building blocks: words, syllables, half-syllables, phonemes, diphones, or triphones [58], [59]. These prerecorded units are concatenated to produce the utterance of interest. However, concatenative synthesis required a lot more data to scale for diverse vocabularies and different speakers.

All these downsides led to the adoption of a learning paradigm in the name of statistical parametric speech synthesis (SPSS) [60], [61]. SPSS adopts the three-stage model presented in Fig. 3, namely, the use of text analysis to suitable linguistic representations of the target utterance, the prediction of speech parameters using an acoustic model, and the final waveform synthesis (vocoding). In particular, the *text analysis* module includes necessary preprocessing steps (text normalization, grapheme-to-phoneme conversion, and so on) followed by the extraction of all relevant features, such as phonemes, duration, or part-of-speech tags. Those features, along with the accompanying speech parameters, are fed to a statistical machine learning (ML) model that learns a mapping from linguistic to acoustic features (e.g., the fundamental frequency, spectrum, or cepstrum). Due to the sequential nature of this data, hidden Markov models (HMMs) have exceled at this type of modeling [60]. Finally, the acoustic features are propagated to a suitable vocoder for the synthesis step. Some notable vocoders are WORLD [62] and STRAIGHT [63]. It is important to emphasize that several (even all) of those steps are *learnable from data*—which is precisely what gave this family of methods its name. Specifically, to learn any of the mappings from graphemes to phonemes to acoustics to the waveform, matching data (i.e., matching text and speech pairs, often obtained from several speakers and of high amount) is needed. This fundamental attribute of SPSS is what makes it the forefather of modern-day deep speech synthesis methods.

B. Deep Speech Synthesis

Deep neural network-based synthesis co-opts neural networks as the models of choice to substitute one or more components of a traditional SPSS pipeline. First attempts usually centered around substituting HMMs with sequential models (RNNs [64] or long short-term memory networks (LSTMs) [65]) for acoustic modeling, such as the early DeepVoice systems [66], [67]. WaveNet was the first neural model to directly generate the waveform from linguistic features [14]. This was later followed by models trying to go directly from character/phoneme sequences to audio, such as Tacotron [68], [69]. Nevertheless, several DL-based methods are still using the traditional three-step pipeline but substitute intermediate steps with their DL counterparts. The defining characteristics of deep speech synthesis are, thus, threefold: 1) methods follow the SPSS formulation; 2) all methods utilize deep neural networks in some, or all, steps of their pipeline; and 3) some methods subsume some or all of the intermediate step in a single model—these are the so-called end-to-end (E2E) approaches.

DL-based methods can be taxonomized along several categories:

1) **autoregressive** (AR) [14] versus **non-AR** (NAR) structures [70];
2) type of **network structure**, where we primarily differentiated between convolutional neural networks (CNNs) [66], [67], sequential models (RNNs, gated recurrent unit networks (GRUs), and LSTMs) [68], [69], which may or may not include attention, and self-attention models (i.e., Transformers) [71], [72], [73];
3) type of **generative model** (e.g., variational autoencoder (VAE) [74] and GAN [75]);
4) **degree of E2E** behavior, which is characterized by the steps of the traditional SPSS pipeline that one or more (jointly trained) models subsume.

While such a categorization is useful for differentiating between different TTS approaches, and later on understanding ESS ones, it is important to stress that the boundaries between those categories are fluid and constantly changing. For example, while WaveNet was first introduced as an AR model, which generates a waveform directly from linguistic features [14], thus integrating the acoustic model and vocoding aspects of an SPSS pipeline, it was later extended to NAR synthesis [76] and changed to produce speech conditioned on (Mel-)spectrograms rather than linguistic features [69]. These rapid changes are expected as researchers continually optimize their pipelines in their quest for E2E synthesis. Nevertheless, as our focus is on presenting the core ideas that have revolutionized the TTS field in the last decade, we will primarily categorize approaches based on their earliest iterations.

An overview of recent, key TTS contributions from the deep learning era is shown in Fig. 2. As previously mentioned, WaveNet [14] was the first neural model to be proposed for speech synthesis. In its first introduction, it was conceptualized as a mapping from textual and prosodic features to a raw waveform—thus integrating the last two steps of an SPSS pipeline. WaveNet also introduced two key innovations in the field of audio modeling: 1) the use of dilated convolutions, which allowed it to increase its receptive field and model long-range interactions and 2) the ability to globally and locally condition the generation process, which proved instrumental in controllable TTS, as well as emotional TTS and voice conversion. Follow-up iterations adapted the model to accept (Mel) spectrograms as input [69], thus effectively transforming it into a more traditional vocoder.

Tacotron [68] approached neural TTS by combining the two frontends of the SPSS pipeline, text analysis, and acoustic modeling, using an encoder-attention-decoder framework. By relying on seq2seq models, optionally augmented with attention, Tacotron learns a mapping from phonemes/characters to spectrograms. These spectrograms are then fed into a suitable vocoder; for that purpose, Tacotron1 used Griffin-Lim [77], whereas Tacotron2 used WaveNet [69]. Due to the sequential nature of the encoder and the decoder, the Tacotron series suffers from slower processing times and difficulties in addressing long-range dependencies.

Following the recent successes of self-attention architectures in modeling such dependencies [78] and their ability to generate their output in NAR fashion by processing their inputs in parallel, Transformers were introduced as an alternative to RNNs in the FastSpeech series [72], [73]. FastSpeech relies on a series of Transformer blocks for encoding the input text sequence; another series of blocks decodes it to the output acoustic features that then serve as input to a suitable vocoder. While Transformers have the advantage of processing the entire sequence in parallel, thus reducing runtime during inference, they require some adaptations to handle the problem of mismatched sequence lengths, as target acoustic features typically have a much longer duration than the input text. This is handled by a duration prediction network, which is trained to predict this mismatch and upsample the learned representations of the encoder to the necessary length before propagating them to the decoder. FastSpeech2 is also trained to jointly predict the pitch and energy of the target speech, which is then used to further modulate the learned representations of the encoder during inference and improve expressivity; this “variance adaptation” mechanism can be readily co-opted for ESS by using it to inject emotional information as well.

Finally, another key contribution to the zoo of neural TTS approaches is the introduction of generative adversarial networks (GANs). Following the seminal work of

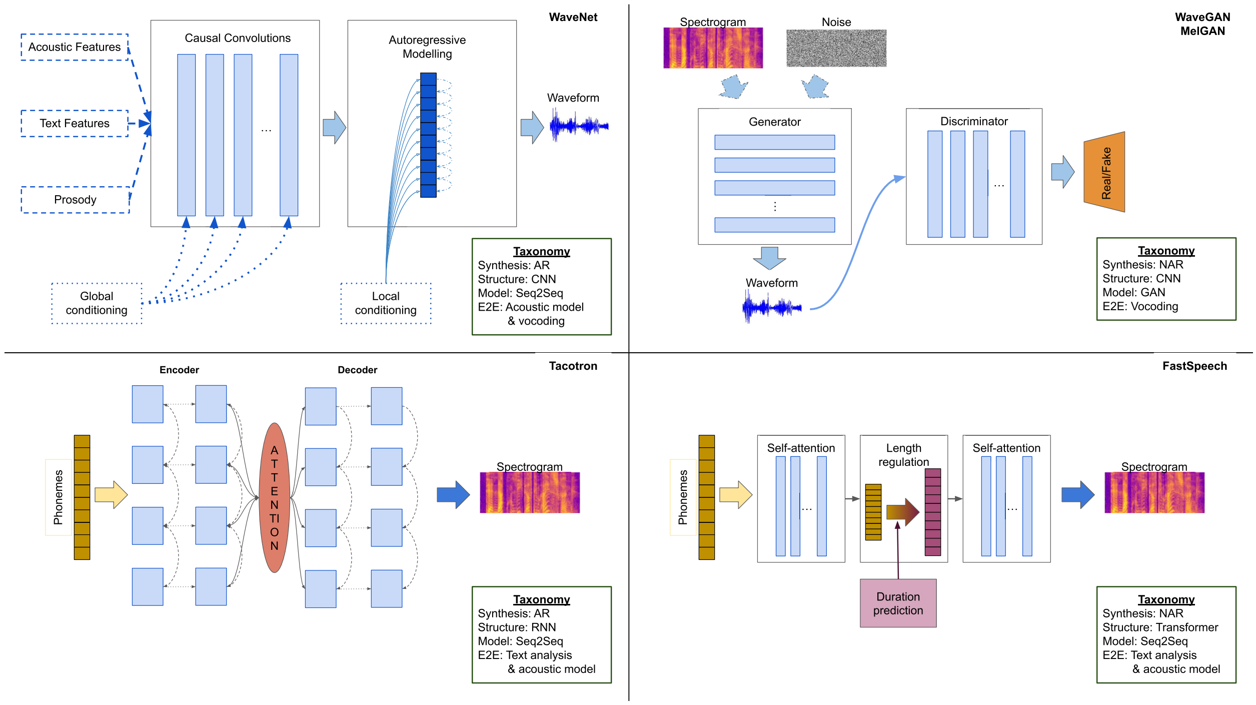


Fig. 2. ***Overview of main deep TTS paradigms. WaveNet was first introduced as a text-to-waveform model (thus combining an acoustic model with vocoding), which could be locally and globally conditioned on additional information; it was later extended to synthesize waveforms from input spectrograms, thus relegated to the role of a traditional vocoder. generative adversarial networks (GANs) are commonly used to map spectrograms to waveforms (effectively acting as vocoders), or to "imagine" waveforms from a random input, as such subsuming all intermediate steps of a TTS pipeline and the mechanism to decide what text to output. Tacotron utilizes sequence-to-sequence (seq2seq) models to learn a mapping from phonemes/characters to audio features, thus implicitly combining text analysis with an acoustic model; FastSpeech iterated on that by substituting recurrent neural networks (RNNs) with Transformers.***

Goodfellow et al. [75], GANs have become mainstays in image, video, and audio generation. For TTS, there are two main categories of GANs. The first one is GAN vocoders, whose generators accept as input spectrograms and output the raw waveform, with the waveform subsequently probed via the discriminator for its "realness." Key examples of this category are MelGAN [79], Parallel WaveGAN [80], HiFi-GAN [81], and others.

Sticking closer to the original formulation by Goodfellow et al. [75], the second category includes models such as WaveGAN [82], which attempt to generate realistic speech from random inputs. They, thus, effectively substitute the entire speech generation model—including the selection of the appropriate text to output—with a single model. While such methods are certainly intriguing, their opaqueness and lack of controllability make them unsuitable for current TTS needs; still, it is an interesting avenue to explore in the search for models that can decide for themselves what they want to say.

We end Section III with a note that we have omitted several key advances in deep speech synthesis. As our goal was not to provide a comprehensive overview, but, instead, a brief one of core novelties introduced in the deep learning era, we have focused on those most pertinent to emotional synthesis. Thus, among others, we have excluded flow- [83] and diffusion-based models [84]. For a thorough review of neural speech synthesis, which also includes these advances, we refer the reader to Tan et al. [15].

C. Deep Voice Conversion

The Voice conversion (VC) is the task of making a speech utterance from a *source* speaker sound like it came from a *target* speaker while keeping the linguistic content unchanged. To further differentiate it from EVC, we also require VC to leave the emotion of the utterance unchanged. Since VC and EVC share many commonalities, we defer a thorough consideration of speech conversion methods to Section IV. Nevertheless, we provide a short synopsis of VC methods here, as it is a vibrant subfield of speech synthesis, with approaches first introduced there and later applied to EVC and vice versa. Two recent comprehensive reviews of VC can be found in [85] and [86].

Several attributes should be manipulated to make the speech of one individual sound like that of another. The first one is the choice of words themselves. Different people use different vocabularies and styles of speaking [86], [87]; therefore, to effectively transform the "identity" of a speech utterance, one should begin with the words that constitute it. However, as we will later also ignore changes to vocabulary necessitated by changes in emotion, we also ignore this important aspect of voice conversion as well. Instead, we focus on the other two attributes: suprasegmental features, such as prosody, and segmental ones, such as spectrum and formants. Short-time spectral features are correlates of *timbre*, which captures the "tone" of an utterance and is highly related to the physiological characteristics of the speaker [87]—though it is also affected by affect and phonetic content. Prosody also captures both physiological characteristics and speaking style [87]. For this reason, several VC works consider only timbre; this, however, limits the success of those methods as human impersonators are found to adapt their prosody as well [87].

VC research has a history of more than 30 years. Early approaches utilized articulatory synthesis but synthesized speech using the parameters of the target speaker [88]. More recent attempts used Gaussian mixture models (GMMs) [89], exemplar-based frameworks based on nonnegative matrix factorization (NMF) [90], and HMMs [91]. However, despite several attempts, no notable progress was made until recent years, which saw the advent of DNNs. Recently proposed methods exploit the representation power of DNNs by means of VAEs [92], [93], GANs [94], [95], and seq2seq models [96], [97].

A key distinction of VC approaches is between those who use *parallel* and *nonparallel* training data. While this distinction will be further elucidated in our discussion of EVC, where it plays an equally crucial role, we already need to touch upon it here. Parallel data mean that utterances of identical linguistic content are available from both the source and the target speaker. While this type of data makes it easier to learn a mapping of the features that capture speaker identity while keeping the content unchanged, they are harder to procure in sufficient quantities, especially for the more data-hungry DL methods. For this reason, algorithms relying on more nonparallel data have become more prominent in recent years.

In general, the power of DL comes in its ability to learn complicated mapping functions from data. In voice conversion, this ability is used to learn a transformation from an input speech signal to the target, usually by transforming the features of the source speaker to those of the target speaker before vocoding. In deep voice conversion, this mapping can be achieved through a conditioning mechanism like the one introduced by WaveNet [14]. As discussed in Section III-B, WaveNet supports both global and local conditioning—these conditioning interfaces can be co-opted by voice conversion algorithms to change suprasegmental and segmental attributes, respectively. The representation of speaker identity, thus, becomes an important aspect of VC. This can be done either by one-hot encodings of a fixed set of speakers [93], d-vectors [92], or bottleneck features as speaker representations from a DNN [98].

As with TTS, GANs also play a prominent role in deep voice conversions. A GAN-VC framework is formulated by using the generator to map an utterance from the source to the target speaker, with the discriminator used to guide the training by classifying whether the target speaker is indeed the correct one. As this mapping can be difficult to learn from nonparallel data, an additional form of regularization is proposed by the introduction of a cycle-consistency loss [99], resulting in CycleGAN [95], [100]. CycleGAN has two generators: one for transforming the speech of the source speaker to the target one and one for the inverse conversion. This is utilized to map the speech of the source/target speaker to the target/source one and back, and ensure consistency (via the L1 loss) with the original source/target utterance. In the process, this ensures that the wanted generator (source to target) is properly trained. An extension of cycle-consistent generative adversarial network (CycleGAN) for multiple speakers is found in StarGAN [94], [101], [102], which extends the consistent principle to multiple source-target domains. Finally, seq2seq models, which use an encoder–decoder architecture, have also been extensively studied in the VC field [96], [97]. These models have the added benefit of handling changes in sequence length induced by changes in features. For example, a change in prosody can make the utterance of the target speaker shorter or longer than that of the source speaker, which cannot be easily handled by frame-to-frame mapping methods.

IV. EMOTIONAL SPEECH SYNTHESIS

In Section IV, we provide an overview of the different DL methods found in recent literature on emotional speech synthesis (ESS). We focus on ESS, rather than broader affective speech synthesis, because these approaches dominated our structured literature search (cf. Section I). While this covers only one particular aspect of affect, we believe that the approaches taken here can inform the synthesis of other speaker traits and traits, and are, thus, representative of the broader field of deep affective speech synthesis. Indeed, the technical mechanisms used to induce expressivity are almost identical to those used to induce emotionality, albeit the latter is constrained to one particular aspect of affect.

ESS is that specific module of an affective agent, which incorporates emotional information in speech utterances by controlling those aspects of speech, which are mostly—ideally exclusively—related to emotion. In its broadest sense, ESS would include a TTS subcomponent, as generating emotional speech does in fact entail the generation of "normal" speech, meaning utterances that are comprehensible both for their intended linguistic and *paralinguistic* meaning. However, most contemporary works envision ESS as an extension of TTS. This is motivated by two pragmatic reasons. First, TTS is a more "basic" problem than ESS, as being unable to procure a comprehensible utterance would make any emotional fluctuations applied to it utterly meaningless. Second, as a corollary of that fact, the research efforts placed on TTS vastly outmatch those placed on ESS. As a result, most approaches are tailored to the former, and the latter is left as a mere afterthought. For these two reasons, ESS approaches mostly rely on an emotional voice conversion (EVC) paradigm, which, like VC, consists of modifications applied to an existing TTS model to control its emotion. Accordingly, the majority of our review will focus on such efforts. We show an overview of the standard ESS pipeline in Section IV-A. As for TTS, we continue with a brief history of earlier works.

A. Emotional Speech Synthesis Pipeline

For present purposes, we focus on the last step of an affective speech generation agent, as discussed in Section II-B, which is the synthesis of the speech utterance itself, after a suitable semantic and affective response have been determined by other processes [36]. Here, though, we place an emphasis on emotion. An overview of this process is shown in Fig. 3, where we present the common blocks of a TTS system and the adaptations required to enrich it with emotion. In brief, a TTS system (blue boxes and lines) comprises three steps: 1) a text analysis module that converts the input text to appropriate linguitic features; 2) an acoustic model that converts those features to acoustic features; and 3) a vocoder, which generates the final utterance. While this is the traditional approach to TTS, the barriers between the different steps have begun to erode with the advent of DL, with a single architecture often subsuming several (or even all) steps.

Incorporating emotionality into this pipeline is primarily done in two ways (green boxes and lines): either an EVC module is tasked with adapting the emotion of the synthesized speech, or the transformation is made as an intermediate step before vocoding. Due to the recent success of TTS architectures, most ESS works are actually performing EVC; however, there are several works that go directly from phoneme sequences to acoustic features, thus incorporating the first two steps of a synthesis pipeline. All these methods will be reviewed in Section IV. Naturally, the target affect may influence the generation of the text response itself, but, as previously mentioned, we ignore this step for our purposes.

It is worth noting that a large portion of the works identified in our literature review has focused on acted, categorical emotions. These more “prototypical” cases are naturally only useful for specific application domains (e.g., movie production); this has also informed how the reviewed methods control the synthesis of emotion and how they evaluate their methods.

B. Brief History of Emotional Speech Synthesis

Cahn’s Affect Editor [8], [9] and Murray’s HAMLET [7], [12] represent the first approaches to emotional speech synthesis. They were both rule-based and relied on the modulation of acoustic correlates of emotion (primarily pitch and timing) and providing those to existing speech synthesizers that took care of the synthesis. The values of these parameters were usually chosen by experts, and their suitability was verified by follow-up recognition studies. A more data-driven study by Burkhardt and Sendlmeier [103] investigated instead different parameter ranges and identified those that lead to better recognition rates, rather than setting them a priori, but, nevertheless, relied on custom manipulating these parameters for synthesis.

While rule-based ESS brought some initial excitement to the field, it was later abandoned in favor of concatenative synthesis [16], [104], [105]. Like its TTS counterpart, concatenative ESS relied on selecting speech units uttered with the appropriate emotions from an existing dataset [29]. As a result, it too suffered from the same shortcomings that plagued standard speech synthesis, especially considering the fact that different speaking styles must be incorporated in the available databases.

Finally, following a similar trend as TTS, ESS transitioned to a data-driven paradigm with the advent of SPSS [106], [107], which, in turn, formed a predecessor to deep ESS. In this context, ESS was primarily envisioned as an intervention on acoustic features before the vocoding step: those features would be mapped to their emotional equivalents before being used to synthesize speech. In particular, mappings between both prosodic and spectral features were learned using data [106], [107]. As is the case for voice conversion, this entailed the presence of *parallel data* from whence the mappings can be learned.

C. Taxonomy of Deep ESS Approaches

There are various ways in that to taxonomize deep ESS approaches, as shown in Fig. 4. The first one is based on whether they perform **TTEFs synthesis**, where they go directly from text to emotionally laden acoustic features, or **emotional voice conversion (EVC)**, where they rely on acoustic features that are already generated by a standard TTS system. Due to the widespread success of TTS, most ESS systems are essentially performing EVC, as they utilize existing components that have proven to work well. One could argue that EVC takes a “shortcut” compared to TTEF since it endows an existing utterance with emotional intonation, rather than synthesizing one from the ground up. This allows EVC to be added as an extra step in a traditional TTS pipeline—first synthesize and then convert to the target emotion. This decomposition into two constituents can reduce computational complexity and dependence on data. Furthermore, vocoding is, to the best of our knowledge, rarely explicitly adapted for ESS, but, instead, existing TTS vocoders are used out-of-the-box.[4] However, as the borders between discrete steps of a traditional SPSS pipeline are eroding with the advent of deep learning, the distinctions between these two forms of ESS are also blurring. Still, as most existing works go under the auspices of EVC, these will form the majority of our review.

EVC methods can be broken down into further categories based on the type of data that they require for training. Approaches relying on **parallel data** learn a mapping

[4] One exception is [129], which uses the emotion label to condition a HiFi-GAN vocoder.

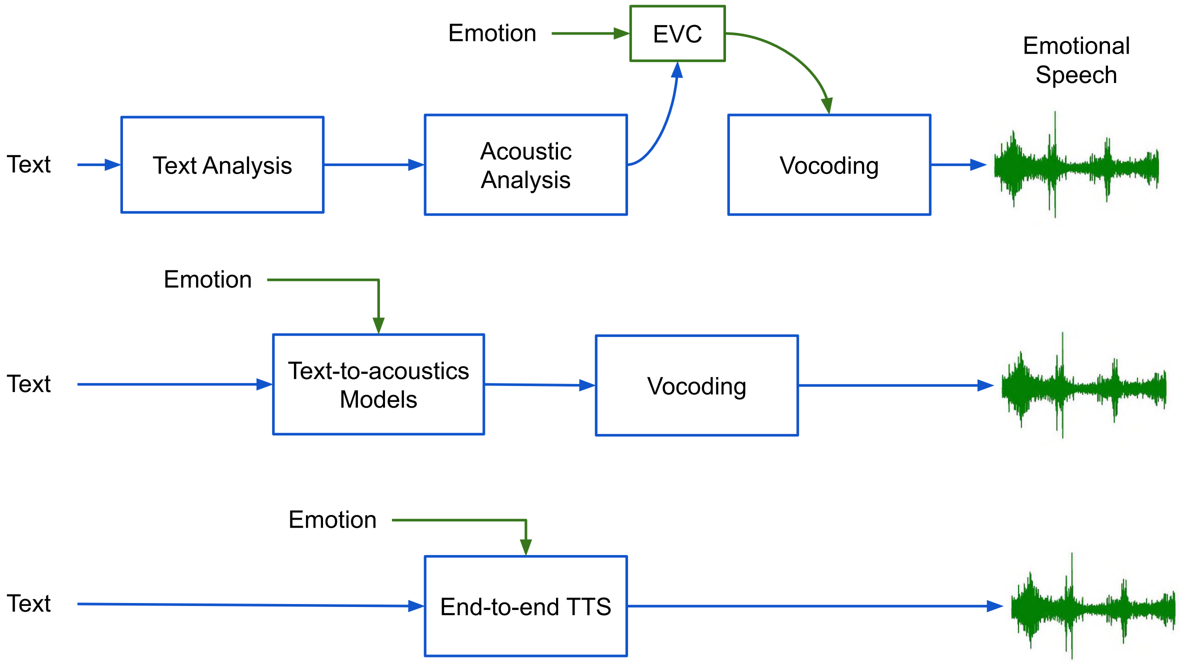


Fig. 3. ***Overview of an emotional speech synthesis module. Emotional synthesis (green) is superimposed on TTS pipelines (blue), which traditionally consists of three steps (top): text analysis, acoustic analysis, and vocoding (synthesis). In the standard setup, emotion is used to modulate the acoustic features before vocoding (emotional voice conversion (EVC)). Deep learning models typically incorporate two (middle) or even all (bottom) of these steps in a single model. In this case, emotion is used as extra, conditioning information to inform the generation of the respective outputs of each model.***

from neutral to emotional speech (or from one emotion to another) by keeping all other factors, such as the speaker or the content, constant; as this approach cannot scale well due to its strict requirements, we only touch upon them in brief. Instead, we focus more on approaches that can work on **nonparallel data**; these scale better as the data can be pooled from several heterogeneous sources. However, they are also more challenging, as the crucial problem of *disentanglement* arises. This challenge gives rise to the second differentiating factor: how the mapping is performed. Some approaches choose a **direct transformation** of one type of speech to another; others opt for a **decomposition** of a speech utterance into discrete components—emotions are one of them, and thus, synthesis can be controlled by choosing one emotional "style" over another.

How emotions are represented to achieve this control is another aspect that we take into account. Here, we differentiate between **reference-based** and **reference-free** approaches. The first kind uses an emotional speech sample to condition an ESS system to the emotion that it needs to produce. The latter provides instead a nonauditory representation of emotions, with the choice of representation being a subcategory of differentiation. Most previous works focus on a limited set of categorical emotions; they, therefore, encode them in "one-hot" vectors ("one-hot" vectors are essentially dummy variables, which transform categorical labels to a numeric representation by using a vector of dimension equal to the number of categories and setting only one of its elements to 1 to represent each category); some works also rely on "fixed" setups, where different DNNs are trained for each one-to-one mapping targeted by the ESS system. Reference-based methods aim for more fine-grained control: this can be achieved by transforming emotional labels to a representation space (usually learned by DNNs) that can be used to increase the span of covered emotions. This is also related to the type of **features** that are manipulated to achieve emotionality: while most choose to modify spectral features, some also use prosody; this, in turn, informs the type of vocoder, which usually follows the feature conversion as it has to support the explicit control of features that are modified. The types of features and control are also dependent on the desired level of **granularity**; utterance-level control is easier to achieve using embeddings, but frame- or word-level controls require more fine-grained approaches.

Finally, as for TTS systems, there are different degrees in which approaches the transition from a multistep (or "cascade") SPSS paradigm, where any one step can be implemented via a DNN, toward an E2E architecture, which incorporates all steps in a single model. This **degree of "end-to-endedness"** will form our last main differentiating factor for ESS approaches. Naturally, as these methods typically modify an existing TTS pipeline, they also

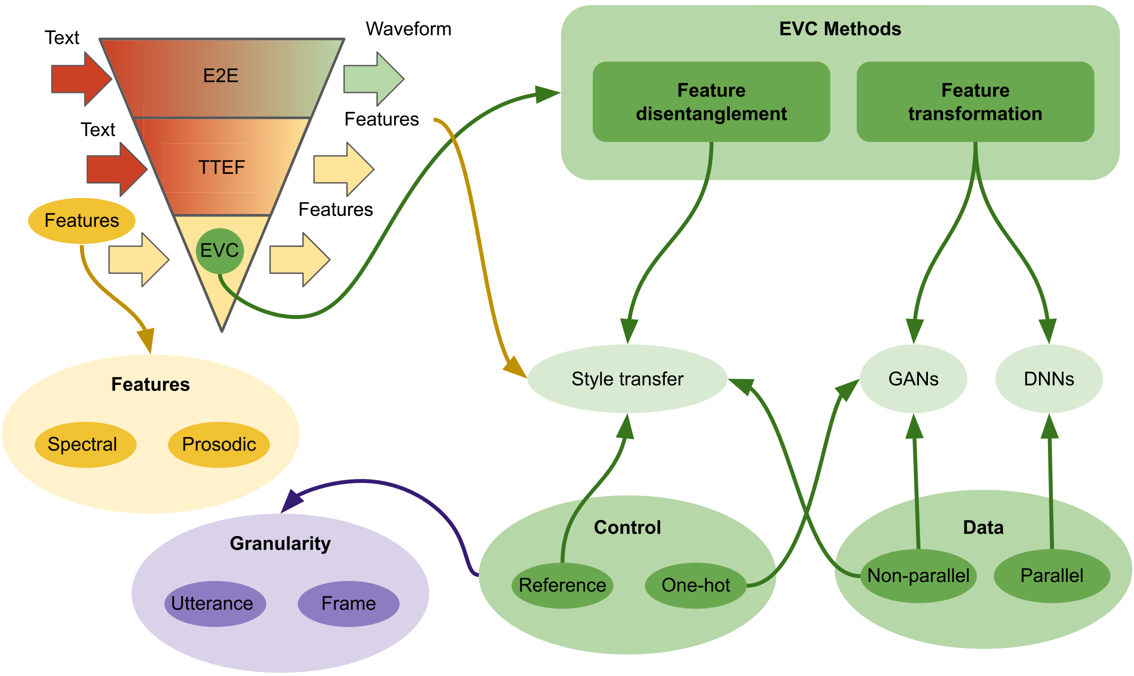


Fig. 4. ***Taxonomy of deep emotional speech synthesis approaches. Approaches can be primarily differentiated according to the following ways: 1) how many steps of the synthesis they incorporate, which is in terms determined by their input and output, accordingly categorized as end-to-end (E2E), text-to-emotional-features (TTEF), or EVC methods; 2) how control is achieved, as well as the level of granularity that this control can achieve; 3) for EVC methods, on whether they use parallel or nonparallel data; and 4) for nonparallel data EVC methods, based on whether they rely on disentangling speech components or directly mapping features to capture the target emotion, as all parallel data methods use the latter form of conversion; TTEF methods, instead, primarily fall under the style-transfer category.***

inherit all its properties, such as the type of architecture or underlying model, as discussed in Section III.

Out of all identified works, we present a selection representative of our taxonomy in Table 1, so as to give a quick overview. Whenever researchers experimented with more than one disjunct category in their work, we chose to assign them according to where their major focus lay.

D. Parallel Data Methods

Parallel data approaches rely on datasets where the same speaker(s) have recorded a set of sentences by acting the entire set of different emotions [110], [113], [120], [136]—similar to parallel VC approaches. This simplifies the conversion problem by keeping all other factors constant; the only thing that needs to be converted is the emotion itself. As such, parallel data methods fall exclusively under the direct transformation category, where a DNN is utilized to learn the mapping between acoustic features or even directly from the text and acoustic features.

However, a major downside of such methods is that they fail to scale, as collecting datasets of sufficient size is difficult given the strict requirements. Moreover, they lack in terms of controllability. As the mapping is dependent on the existence of parallel data, it is only possible to map from one emotion, or from neutral speech, to another type of emotion, and this mapping is fixed. Often, it is the case that researchers train distinct networks for each combination in their dataset.

Finally, parallel datasets are often recorded in very controlled conditions. Usually, a single speaker or a small set of speakers record a small set of sentences in one room using the same microphone and are acting the required emotions. This vastly differs from real-world situations where emotions have to be naturalistic and fit a number of different environments. Therefore, parallel data methods cannot generalize well in the scenarios expected for real-world applications.

In conclusion, while simplifying the conversion problem by fixing other variables offers several advantages, primarily via simplifying the underlying problem, the downsides limit the applicability of the developed ESS systems. Thus, parallel data methods were mostly pursued in the early days of deep ESS as a means of prototyping.

E. Nonparallel Data Methods

Transitioning to methods capable of handling nonparallel data was a necessary prerequisite for the development of more naturalistic and generalizable ESS systems. This enabled the use of larger datasets, often ones used in SER research, collected in less controlled conditions.

This transition, though, introduced a major challenge: several factors were now entangled in the utterances to

Table 1 Major DL-Based Emotional Speech Synthesis Works Categorized According to the Taxonomy Presented in Fig. 4

Approach	Control	Intensity	Granularity	Non-parallel data	Conversion	Model	Features	End-to-end
Ming *et al.* [110]	Fixed	N/A	Utterance	✗	Transformation	bLSTM	STRAIGHT	EVC
Lee *et al.* [111]	One-hot	N/A	Utterance	✗	N/A	Seq2Seq	Spectra	TTEF
Lorenzo-Trueba *et al.* [112]	Annotator agreement	Annotator agreement	Utterance	✗	N/A	RNN	WORLD	TTEF
An *et al.* [113]	Reference	N/A	Utterance	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Choi *et al.* [114]	Reference	N/A	Utterance	✗	Disentanglement	CNN	Spectra	TTEF
Kwon *et al.* [115]	Reference	N/A	Utterance	✗	Disentanglement	Seq2Seq	Spectra	TTEF
Shankar *et al.* [116]	Fixed	N/A	Utterance	✗	Transformation	Highway	F0/intensity	EVC
Bao *et al.* [117]	Fixed	N/A	Utterance	✓	Transformation	CycleGAN	openSMILE	EVC
Luo *et al.* [118]	Fixed	N/A	Utterance	✗	Transformation	GAN	F0	EVC
Robinson *et al.* [119]	Fixed	N/A	Frame	✗	Transformation	Seq2Seq	F0	EVC
Gao *et al.* [120]	Reference	N/A	Utterance	✓	Disentanglement	GAN	F0/Spectra	EVC
Aggarwal *et al.* [121]	Reference	N/A	Utterance	✓	Transformation	Seq2Seq	Spectra	TTEF
Kim *et al.* [122]	Reference	N/A	Utterance	✗	Disentanglement	Seq2Seq	Spectra	EVC
Rizos *et al.* [123]	One-hot	N/A	Utterance	✓	Transformation	StarGAN	Cepstra	EVC
Cao *et al.* [124]	Fixed	N/A	Utterance	✓	Transformation	VAE-GAN	Cepstra	EVC
Lei *et al.* [125]	Reference/Fixed	Ranking	Phoneme	✓	Transformation	Seq2Seq	Cepstra	TTEF
Schnell and Garner [126]	Reference	Saliency maps	Frame	✗	Transformation	RNN	WORLD	EVC
Liu *et al.* [127]	Reference	N/A	Utterance	✓	Transformation	Seq2Seq	Spectra	TTEF
Du *et al.* [128]	Reference	N/A	Utterance	✓	Transformation	StarGAN	Cepstra	EVC
Choi and Hahn [50]	Reference	Manual	Utterance	✗	Disentanglement	Seq2Seq	Spectra	EVC
Cai *et al.* [129]	Reference	N/A	Utterance	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Wu *et al.* [130]	Reference	N/A	Frame	✗	Disentanglement	Seq2Seq	Spectra	TTEF
Kreuk *et al.* [131]	Fixed	N/A	Frame	✗	Transformation	Seq2Seq	Spectra/F0/T	EVC
Chen *et al.* [132]	Reference	N/A	Utterance	✗	Disentanglement	CNN	Spectra	TTEF
Zhou *et al.* [51]	Reference	Ranking	Utterance	✗	Disentanglement	Seq2Seq	Spectra	EVC
Zhang *et al.* [133]	Reference	Posterior	Utterance	✓	Disentanglement	Seq2Seq	Spectra/F0	EVC
Li *et al.* [134]	Reference	Manual	Utterance	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Liu *et al.* [135]	Reference	N/A	Frame	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Lei *et al.* [136]	Reference	Ranking	Frame	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Feng *et al.* [137]	Sentiment	N/A	Utterance	✗	Transformation	Seq2Seq	Spectra	TTEF

be processed. In particular, there was now no matching sentence of the target emotion for the text that needed to be synthesized (or converted), and sometimes, the data for the target emotion even came from a different speaker. This necessitated the disentanglement of those different factors. This entails the decomposition of an input utterance to a set of independent constituents, the modification of the emotional style (and, if needed, the speaker identity), and the reconstruction of the resulting waveform.

As we show in Sections IV-F and IV-G, this decomposition could be implicitly enforced to the model via manipulation of its training strategy, leading to a set of methods that we name *direct feature transformation methods*, or explicitly designed into it, leading to *disentanglement methods*. In either case, the increased complexity of handling nonparallel data, followed by the concurrent advancement of TTS systems, resulted in most researchers adopting an EVC paradigm. They resorted to modifying the acoustic features to induce emotionality and relied on existing TTS pipelines for all other aspects of the synthesis process. Only recently did they transition back to TTEF approaches, following the success of Tacotron and similar TTS models. This trend is also reflected in Table 1.

F. Feature Transformation Methods

The first category of EVC methods attempts to learn a direct transformation between features of one emotion to another. In its simplest form, the EVC problem can be formulated as follows. Given a set of input features $X_S \in \mathbb{R}^{T_s \times d}$, with T_S being the duration of the utterance and d the dimensionality of the features, the goal of an EVC model is to map those to $X_T \in \mathbb{R}^{T_t \times d}$, which are the target features of a potentially different duration but of the same dimensionality. This mapping takes the form of a function $f(\cdot) : X_S \rightarrow X_T$, which will be approximated by a DNN as $\tilde{f}(\cdot)$. Concretely

$$x_t = f\left(x_s\right) \approx \tilde{f}\left(x_s\right) \tag{1}$$

where x_s and x_t are samples sourced from X_S and X_T, respectively. As we saw in Section IV-D, this mapping is easier to learn when X_T and X_S have the same linguistic content and come from the same speaker; it is then sufficient to train a DNN to estimate the mapping between paired utterances as all this DNN will capture is the change in emotion. However, nonparallel methods that rely on direct transformation still have to deal with the problem of entangled factors.

The most widespread method that deals with this problem is generative adversarial networks (GANs). GANs were first introduced by Goodfellow et al. [75] and follow the basic idea of having two neural networks that compete against each other, hence the name *adversarial*. In the originally proposed GAN framework, one of those networks, the so-called *generator*, learns a function that transforms noise vectors z sampled by a particular random distribution to the data x that follow another distribution, where this target distribution resembles data of a specific target domain. The second network, the *discriminator*, tries to classify those artificially created instances as "fake" data. During training, the discriminator is fed with real data from a given training set and fake data created by the generator. Its objective is to distinguish between these two classes of data. On the contrary, the generator's goal is to "fool" the discriminator by learning to generate realistic data. Thus, the two networks have conflicting goals, resulting in both of them mutually improving each other during training. The combined objective function V of the GAN framework can be formalized as follows:

$$\begin{aligned}\min_{G}\max_{D} V(D,G) &= \mathbb{E}_{x\sim p(x)}\left[logD(x)\right] \\ &+\mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \quad (2)\end{aligned}$$

where G is the generator network and D is the discriminator. Since the input vectors z were sampled from a random probability distribution function, the generation of completely new data is possible by merely feeding different random vectors into the generator.

An intuitive extension of this principle for EVC would be to replace the random noise vectors with data that follow the distribution of a certain speech type, i.e., a dataset of natural speech, while using the output of the generator and a dataset of the target speech type as input for the discriminator. This would lead to the generator transforming the input speech of a certain source domain to the speech of the respective target domain. However, these kinds of translation approaches only work well with paired training data, as, otherwise, the discriminator would easily detect the distributional changes induced by differences in the speaker or linguistic content. As a consequence, the standard GAN paradigm needed to be modified for nonparallel training data.

One of the most prominent approaches that deal with the aforementioned problem is the cycle-consistent generative adversarial network (CycleGAN) [99]. CycleGAN combines two unique GANs, each consisting of its own generator and discriminator. The idea is that one GAN learns to transform data from a domain X_S to a domain X_T, whereas the other GAN learns the exact opposite: converting data from domain X_T to domain X_S. Thus, by feeding input data of one domain to one of the GANs, and subsequently feeding the output of that first GAN back into the second one, the final result can be compared with the original input. In the case of both GANs working perfectly, the final result should be exactly the same as the initial input. During training, CycleGAN uses a *cycle-consistency loss* as part of its objective function, in addition to the adversarial loss that is adopted from the original GAN architecture. The cycle-consistency loss is formulated as

$$\begin{aligned}\mathcal{L}_{\text{cycle}} &= \mathbb{E}_{x_s\sim p(X_S)}\left[\left\|G_{T\to S}\left(G_{S\to T}\left(x_s\right)\right)-x_s\right\|_1\right] \\ &+\mathbb{E}_{x_t\sim p(X_T)}\left[\left\|G_{S\to T}\left(G_{T\to S}\left(x_t\right)\right)-x_t\right\|_1\right] \quad (3)\end{aligned}$$

where $G_{S\to T}$ and $G_{T\to S}$ are the generators of the two GANs, and x_s and x_t are data from the domains X_S and X_T, respectively. The full loss function of a CycleGAN is given as

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{GAN}}\left(G_{T\to S}, D_S, X_S, X_T\right) \\ &+\mathcal{L}_{\text{GAN}}\left(G_{S\to T}, D_T, X_T, X_S\right) \\ &+\lambda\mathcal{L}_{\text{cycle}}\left(G_{T\to S}, G_{S\to T}\right) \quad (4)\end{aligned}$$

where D_S and D_T are the discriminators of the two GANs and λ is a balancing factor.

CycleGAN was first transferred to the speech domain by Kaneko and Kameoka [95], who used it to perform voice conversion. In order to do so, they enhanced the generator networks with *gated CNNs* [137]. In addition, they implemented an *identity loss* [138], which ensures that samples already belonging to the target domain are not altered. They formulate the identity loss as

$$\begin{aligned}\mathcal{L}_{\text{identity}} &= \mathbb{E}_{x_t\sim p(X_T)}\left[||G_{S\to T}\left(x_t\right)-x_t||_1\right] \\ &+\mathbb{E}_{x_s\sim p(X_S)}\left[||G_{T\to S}\left(x_s\right)-x_s||_1\right]. \quad (5)\end{aligned}$$

A more sophisticated version of their work that introduces additional discriminator networks that are applied on the circularly converted voice was presented in [100]. Later, EVC approaches then adopted this formulation for their purposes [115], [139], [140], [141].

CycleGAN though is faced with a major limitation, namely, that it only supports a translation between two domains. However, it is desirable that EVC methods cover a wider set of emotions. This would mean separately training an equal set of CycleGANs, which, aside from increasing the computational overhead of experiments, also fails to benefit from the synergistic effects that may arise from a many-to-many mapping. This problem was solved by CycleGAN's successor: StarGAN [101], [121], [126], [142]. The training concepts of both models are illustrated in Fig. 5.

The basic concept of StarGAN is to use a single generator and discriminator, both conditioned on features of multiple domains during training. The conditioning information, i.e., the emotion in the case of emotional speech conversion, is given as a domain code c. The discriminator is fed with this domain code in combination with the input audio, while the generator is conditioned with a different domain code c' representing the target domain. Thus, the adversarial part of StarGAN's objective is formulated

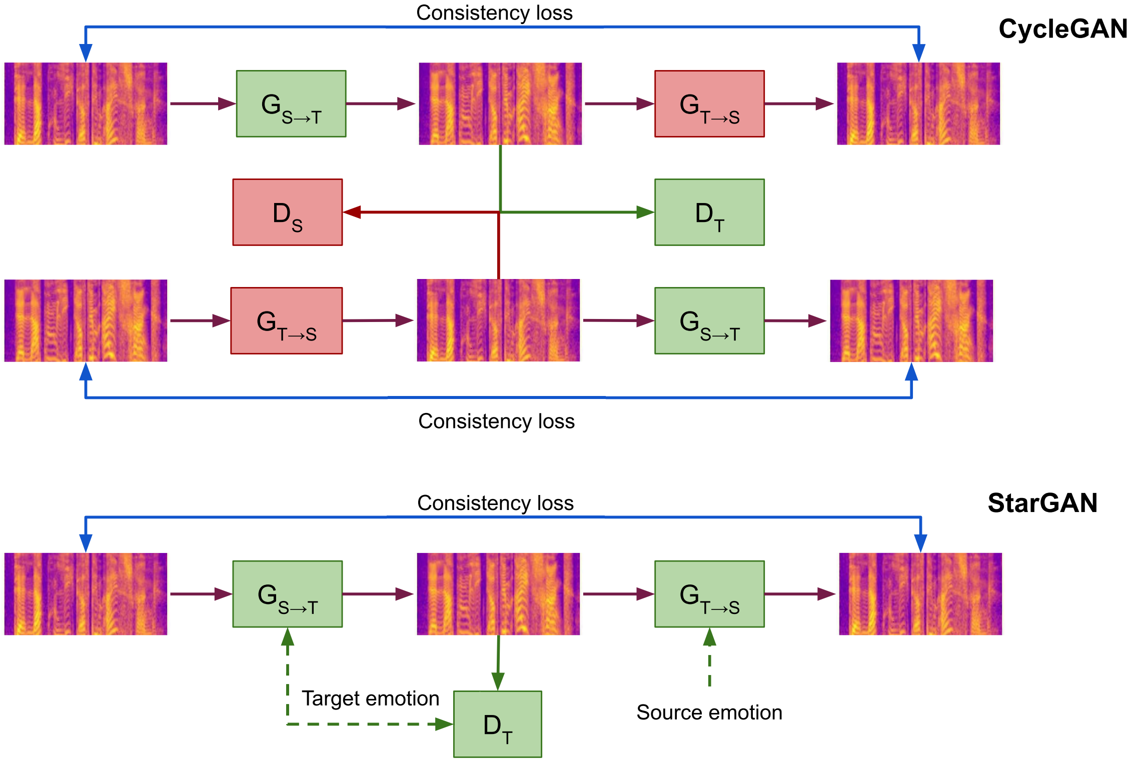


Fig. 5. ***Overview of the two main GAN paradigms for emotional voice conversion. CycleGAN (top) first converts the source utterance to the target domain, which is then evaluated by the discriminator, while another generator maps it back to its source domain to compute the consistency loss. The inverse process is followed to map from the target domain to the source domain, which helps to regularize training. StarGAN (bottom) extends CycleGAN to handle a many-to-many mapping by training a single generator and discriminator pair, both of which can be controlled via one-hot labels to perform the correct mapping/check.***

as follows:

$$\begin{aligned}\mathcal{L}_{t-\text{adv}} = {} & \mathbb{E}_{(x,c)\sim P(x,c)}\left[logD\left(x,c\right)\right] \\ & + \mathbb{E}_{x\sim P(x),c'\sim P(c')}\left[\log\left(1-D\left(G\left(x,c'\right),c'\right)\right)\right].\end{aligned} \tag{6}$$

In addition, to enforce the model to create audio that belongs to the target domain, a classification loss is added. To do so, an auxiliary classifier C is trained alongside the discriminator and generator to distinguish between the different domains. This classifier is used to build StarGAN's classification loss component

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{x\sim P(x),c'\sim P(c')}\left[-\text{log}C\left(c'|G\left(x,c'\right)\right)\right]. \tag{7}$$

Furthermore, analogous to CycleGAN, a cycle-consistency loss and an identity-mapping loss are used.

G. Feature Disentanglement Methods

Disentanglement methods attempt to explicitly decompose into several constituents: emotion, linguistic content, and, potentially, speaker effects. This is tantamount to assuming that the speech signal results from an equivalent number of latent factors, c (for content), e (for emotion), and s (for speaker), which together forms the latent code z defined over a representation space $\mathbb{Z}$. This latent code is mapped to the observation space $X \in \mathbb{R}^{T\times d}$ via a function $f : \mathbb{Z} \to \mathbb{R}^{T\times d}$. The core idea is to preserve the factors related to content and speaker while manipulating the ones related to emotion.

The basic schema, as followed by recent works [50], [127], [131], are shown in Fig. 6. The source utterance features are passed to an encoder whose goal is to learn the content, resulting in an embedding c_s; a reference embedding (or some other representation of emotion) is passed to an emotion encoder, which generates an emotion embedding e_t; and both embeddings are passed to a decoder, which attempts to reconstruct the target utterance. If no voice conversion is required, it is assumed that the content embedding c_s also contains speaker information, which should be preserved; otherwise, a speaker embedding s_t is also created by a speaker encoder. In order to guide the respective encoders to properly disentangle the information that they need, specific losses are introduced. For example, the embeddings of the emotion encoder could be compared to embeddings from an SER model or passed to an SER model themselves such that they learn to classify the target emotion [50], [132], [133], [143]. This way, they are guided to learn emotional information. Finally, another way to disentangle speaker and emotion

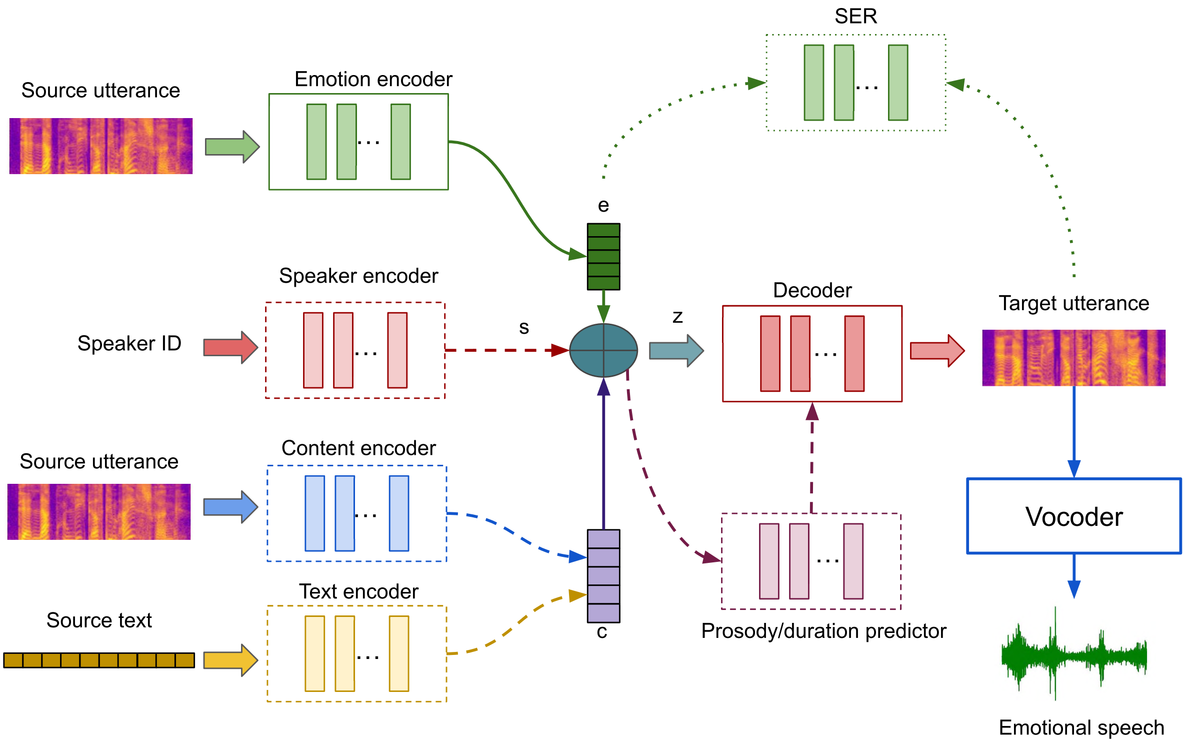


Fig. 6. ***Overview of disentanglement methods for emotional voice conversion. The source utterance is passed through an emotion encoder to generate an emotion embedding. Content information is additionally provided by encoders operating on the acoustic or linguistic features or both. Optionally, speaker embeddings are provided by a speaker encoder. All embeddings are fed to a decoder that generates the target utterance. This utterance and the emotion embeddings are often evaluated with respect to the emotional information that they contain, usually by a pretrained SER model. Additional information, such as prosody or duration, is often predicted from the joined embeddings and propagated to the decoder to improve the quality of synthesis. During inference, the source utterance or text is used to provide the content, while the reference utterance is used to generate an emotion embedding.***

information is to explicitly perturb features that are expected to correlate with emotion (e.g., pitch) and those related to speaker identity (e.g., formants) [144] and rely on a model to learn emotion-specific perturbations that help map to the target emotion; however, this approach is rather limited because both types of features are related with both variables. This limitation could be combined with adversarial approaches that attempt to remove speaker information from prosodic representations [145].

Inspired by advances in TTS, with models such as Tacotron achieving impressive synthesis results, recent works have transitioned to conditioning such models on emotional information [127], [128], [132], [133], [134], thus moving a step up from EVC methods in the E2E hierarchy and going directly from the text to audio features. These methods have no content encoder relying on audio inputs; instead, the text encoder acts as the content encoder to capture linguistic style. While the information fed to the content encoder is already decoupled from its expressivity (as it comes in the form of characters or phonemes), we consider these methods to be relying on disentanglement, as they still rely on an emotion encoder to capture emotional information and a content encoder to capture linguistics. The main difference is that this decomposition is now implicit in the content part.

We note that some works explicitly attempt to disentangle *prosody* and *timbre*, using the former to denote pitch, tempo (local speaking rate), and intensity (loudness), and the latter to stand for all other aspects that characterize sound [131]. This dichotomy implicitly (sometimes explicitly) assumes that prosody—as defined by the three dimensions of pitch, tempo, and intensity—is the only carrier of affective information. However, several aspects of timbre have been shown to function as affective cues. For example, voice quality has been found to vary a lot depending on speaking style, and to do so independently of pitch or intonation [48], [146], while also functioning as a marker of personality [147]. This is also encapsulated in the design of SER feature sets that incorporate voice quality measures as markers related to emotion [148]. Usually, timbre is left as the residual information to be learned by a speaker encoder, whose function is to condition the TTS system on the voice characteristics of the target speaker (using a reference utterance for unknown speakers or a hard-coded speaker embedding for known ones). These models may, thus, inadvertently convey the affective information included in the reference utterance in the form of voice quality or, alternatively, fail to capitalize on the expressive capacities of timbre by not controlling it.

H. Controlling the Synthesized Emotion

Controlling the generated emotion is a fundamental aspect of ESS systems. In fact, this is the most important differentiating factor between methods falling under the ESS umbrella and *expressive TTS* systems, which impart prosodic fluctuation on the generated signals. Without explicit control, these fluctuations influence other aspects of the output utterance besides emotion, such as speaker identity or an overall speaking "style" inherited by the training data (for example, read versus spontaneous speech). While ESS methods make great efforts to properly disentangle these aspects, they also need a mechanism to control the emotion during inference.

Early ESS methods relied on a trivial control mechanism: by independently training conversion systems for all possible emotion pairs, the inference was simply done by selecting the appropriate pair. This is essentially superseded by methods that use one-hot encodings of emotions, such as StarGAN [121]; there each emotion is represented by a one-hot vector [149]. These approaches are better able to handle many-to-many mappings by jointly learning from several emotion pairs.

However, relying on fixed codes is far below the level of control required for successful HCI applications. For this reason, later methods resort to reference-based "style transfer" methods, which are inspired by recent advances in expressive TTS. The core proposition behind style transfer is to learn an encoding for those aspects of speech that corresponds to emotion in a data-driven way. This idea was first introduced in [150]. The authors utilized a reference encoder to capture prosodic variations in the reference signal and transfer them to the target utterance. This encoder was jointly trained with the main TTS system; during training, the reference was identical to the target utterance, so a mapping could be learned; during inference, the reference was chosen arbitrarily to impose a specific style on the target utterance. Their reference encoder was a CNN followed by a GRU that relied on Mel-spectrogram features; the style embedding was simply the last state of the GRU, which was then fed to the decoder. However, according to the authors' experiments, the learned embeddings captured heavily entangled information, which is highly undesirable as it inhibits the fine-grained control of the synthesis process. For example, transferring from a female voice to a male voice resulted in an overall lower pitch, which sounded like a female trying to imitate a deeper voice [150].

To further promote disentanglement between different factors, Wang et al. [151] introduced global style tokens (GSTs). Utilizing a similar setup as in [150], they introduced an additional constraint on the reference encoder. Rather than propagating the last state of the GRU to the decoder, they first use it as the query in attention operation over a set of learnable tokens: the GSTs. These tokens, whose number was fixed a priori to 10, would be the knobs to be twisted during inference time. During training, the attention mechanism would "softly" weigh the contribution of each token to the reference; the contributions would then be combined and fed to the decoder. During inference, the user can either provide a different reference, which would be accordingly run through the reference encoder to generate the GST weights, or directly manipulate the weights themselves to achieve the required outcome. Inducing this constraint resulted in more naturalistic control and better disentanglement in the styles.

Both these procedures are often used for EVC. The main difference is in the type of utterances provided to the reference encoder. During inference, these are selected to belong to the target emotion, similar to the expressive TTS case where a reference is picked to fit the required style. However, in EVC, it is also to explicitly guide the tokens to encode the reference emotion during training [127], [132], [133], [152], [153]. This already biases the reference encoder to capture the differences in emotion and results in better controllability during inference as well. Overall, this draws inspiration from image style transfer; in fact, in one particular instance, style transfer has been explicitly used on Mel-spectrograms that have been converted to images [154]. Moreover, it is often the case that an average style is computed on all training data and stored as a reference for each category, therefore sparing the need to select an appropriate reference during inference though some works attempt to derive more representative styles than simply averaging the context vectors [155], [156].

Finally, a particular line of work uses text sentiment to control the synthesized emotion [135], [157] or its strength [123]. This draws inspiration from Stanton et al. [158], who directly predict GSTs from text and, thus, do not require an external reference utterance. For some applications, this intermodal consistency is important; for others, though, it may be too limiting as it does not allow for expressing a different emotion in the audio than in the text (e.g., irony). Thus, despite being an interesting form of control, it is not as widely applicable as the other instances discussed above.

I. Controlling Emotional Intensity

Another vital aspect of control is that of emotional intensity. Emotions come not in discrete states but in continua, which define fluid categories that seamlessly transition from one to the other [53]. As such, controlling the strength of a required emotion—over a continuous axis—is of fundamental importance for ESS. In general, this area has remained relatively underexplored as researchers grapple with the challenges of discrete ESS. Nevertheless, there have been important advances in recent years.

Lorenzo-Trueba et al. [110] were probably the first to investigate it for deep ESS. They used an annotator-driven representation of emotion, which assigned a relative value to each utterance. This relative value was computed using the confusion between the expected emotion (i.e., what the actors in their dataset were supposed to act) and the annotated emotion (i.e., what annotators perceived). This allowed them to represent emotions via continuous, rather than discrete, vectors, which, in turn, allowed for the fine-grained control of emotional synthesis via manually setting those weights during inference. A similarly manual setting was also explored by Choi and Hahn [49] and Li et al. [132].

Being simple in its conception, this form of control is fundamentally limited by the lack of data or the need to manually tune parameters. An interesting alternative is found by Schnell and Garner [124], who use the saliency maps of pretrained, attention-based SER systems as guidance for frame-level intensity control. Similarly, Zhang et al. [131] use the posterior probabilities of an SER for utterance-level emotional control. Both approaches follow a reference-based paradigm for inference-time control. Ultimately, those approaches are limited by the effectiveness of the SER systems, which, although greatly improving in recent years [159], is still far from perfect. Moreover, it is not necessarily the case that those references evaluated by an SER system as more "probable" are necessarily those of a higher intensity; they could merely be those that are closest to its training distribution.

A solution is given through exploiting the inherently ordinal nature of emotions [160]. Lei et al. [123], [134] and Zhou et al. [50] exploit this fact by learning a ranking function for the intensity of each emotion. Their approach relies on assuming that all neutral samples have an intensity of zero and proceeding to generate emotional-neutral ranking pairs, as well as neutral–neutral and emotional–emotional anchor pairs, over which a max-margin optimization problem is approximated. This results in a weighting matrix W, which provides a ranking between [0–1] for each feature vector x. During inference, this ranking can be manually set to control the intensity of the synthesized emotion.

J. Granularity of Emotional Control

Most works impose a single emotional category or style on an entire utterance, assuming that this will be accordingly "distributed" by the mapping network or decoder to the appropriate frames. However, achieving a more fine-grained level of control can help increase the naturalness of expressed emotions and add the capacity to express more nuanced emotional states. To that end, some works pursue more granular representations of emotion. For example, Schnell and Garner [124] are able to achieve this through their saliency maps, which assign a level of control to each frame via an attention-based SER model. Similarly, Wu et al. [128] achieve this with a capsule network [161], while Kreuk et al. [129], Liu et al. [133], and Lei et al. [134] achieve this via frame-level losses. Lu et al. [162] and Lei et al. [123] both pursue phoneme-level control: the former via learning a speaker- and emotion-independent phoneme latent code (using gradient reversal to remove this information from their reference encoder), while the latter simply maps their utterance-level emotion intensity to phonemes. This latter approach is similar to Liu et al. [20], who do emphatic speech synthesis and linearly interpolate across the phonemes of emphasized words for more granular control. As seen in Table 1, this trend is picking up pace this last year with several very recent works pursuing higher degrees of granularity.

K. Features Manipulated to Achieve Expressivity

As shown in Table 1, most methods fall under the EVC category, meaning that they primarily manipulate acoustic features to achieve expressivity. These features are to a large extent motivated by the decades of research devoted to understanding which facets of speech are impacted by emotion and how. This research is touched upon in Section II-D. Another factor that influenced the choice of features is the success of modern TTS architectures. As we saw, EVC methods heavily rely and often outright incorporate, existing TTS pipelines. It is only natural that they then use the same features that those TTS pipelines support. Previously, this restricted the set of features to ones supported by SPSS vocoders, such as WORLD or STRAIGHT, which included F0, spectral/cepstral, and aperiodicity features [108], [110], [117], [121], [124], while works that only intended to use EVC as a means to improve SER performance [115] manipulated feature vectors used by the downstream models [163]. Nowadays, with the advances seen in neural vocoders, it is typical to use those as the last step of the synthesis process; accordingly, EVC pipelines now concentrate more on modulating spectral features [50], [109], [112], [113], [120], [140]. Manipulation of other features, such as F0, is still done but primarily using simple statistical techniques (e.g., by standardizing the F0 curve with the statistics of the target emotions [121]) and rarely using deep learning methods [114], [129], [164], [165]. Overall, this shows that the field is transitioning to a standard of using more abstract representations (spectrograms) and relying on the representation power of DNNs for learning to modify the appropriate signal characteristics.

L. Deep Models Used in ESS Research

In general, from an architecture perspective, the innovation in the field of ESS does not seem targeted to novel DL models, but rather on finding novel ways of combining existing modules to achieve desired effects (e.g., disentanglement). This is to be expected following the success of TTS; adopting best practices from a neighboring field allows the community to iterate quickly over problems that are specific to ESS rather than reinventing the wheel. As seen from Table 1, the majority of EVC models are relying on seq2seq models [50], [109], [113], [117],

[120], [124], [125], [131], [153]. This is counter to earlier methods, which relied on highway networks or simpler sequential models [108], [110], [112], [114]. The main downside of those was that they could not handle the differences in signal duration that resulted from a change of emotion; thus, this mapping of duration needed to be handled explicitly. In contrast, seq2seq methods have a natural way of handling the change in duration as the decoder can reconstruct sequences of different lengths than those seen by the encoder [18]. A more thorough overview of seq2seq models for EVC can be found in [18]. This seq2seq trend is also followed by more recent TTEF methods, which rely on the Tacotron architecture [127], [128], [132], [133], [134] and merely condition it with emotional information. Methods using adversarial models, such as CycleGAN [115] and StarGAN [121], also stay close to their original versions, with minor adaptations to fit the EVC problem. Finally, some works attempt to leverage representations learned by large, pretrained models and, thus, rely on transformer-based architectures, such as HuBERT [129].

The decomposition of input utterances, either source or reference, is typically achieved via the use of autoencoders (AEs); this makes them foundational building blocks of several EVC methods. They, therefore, warrant a closer analysis. Traditional AEs are comprised of two parts: 1) an encoder, which reduces the dimensions of the speech signal to a latent representation (or code) and 2) a decoder, which tries to reconstruct the original speech representation from the code. Mathematically, given the speech frame of dimension d $\{x \in R^d\}$, we define the encoder as a function $q : \mathcal{X} \rightarrow \mathcal{Z}$ such that $z = q_\phi(x)$ with parameters ϕ, and the decoder as a function: $\psi : \mathcal{Z} \rightarrow \mathcal{X}$, such that $\tilde{x} = \psi_\theta(z)$ with parameters θ. The same principles are used for EVC; however, instead of a single encoder, there are often multiple ones, one for each latent factor that needs to be disentangled, while a single decoder takes care of the inverse mapping to the feature space.

A probabilistic realization of AE that is sometimes used for EVC as well is the VAE [119], [122], [157], [166], [167]. A VAE is used to generate the speech representation of the target domain, where the code of the network is assumed to be represented by a Gaussian distribution $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mu, \mathrm{diag}(\sigma^2))$. The encoder tries to estimate the mean and variance of the distribution, and with the use of the reparameterization trick, we can sample a code representation. The code is fed to the decoder, which estimates a new speech representation. The training is performed by maximizing a variational lower bound of the log-likelihood

$$\mathcal{L}\left(\theta, \phi; x\right) = -D_{\mathrm{KL}}\left(q_\phi\left(z|x\right)||p\left(z\right)\right) + E_{q_\phi}\left[\log \psi_\theta\left(x|z\right)\right] \tag{8}$$

where $D_{\mathrm{KL}}(q||\psi)$ denotes the Kullback–Leibler divergence between the distributions q and p.

M. Evaluation Protocols

Protagoras of Abdera famously claimed that "of all things the measure is Man, of the things that are, that they are, and of the things that are not, that they are not," and so is the case for the evaluation of ESS approaches as well, with the employment of human annotators being the gold standard for judging the effectiveness of ESS approaches. The most commonly used process is a judgment test, where annotators are asked to evaluate the similarity of a generated signal with respect to a reference stimulus or, in the reference-free variant of those tests, to simply classify the emotion of the generated signal. Alternatively, they are asked to evaluate different signals with respect to different aspects that correspond to emotional speech, such as likeability, emotional strength, or naturalness. In all cases, individual ratings are aggregated to procure a final mean opinion score (MOS) [168]. Usually, these ratings are on a [0, 5] scale with steps of 0.5, with 5 being the best score. While there is currently a dearth of well-established dimensions on which to evaluate emotional speech, the field is drawing inspiration from the much more mature metrics for TTS [169]. Some dimensions commonly used in recent works are naturalness [49], [116], [118], [133], speech quality [124], [126], emotional strength [110], and similarity with the target emotion [112], [113], [128], [131], [132]. ABX tests are also commonly used [116], [126], where subjects are asked to tell if sample X, which is randomly chosen from category A or category B, is closer to a reference from A or a reference from B [170]. If subjects systematically pick the correct category for sample X, the two categories are considered to be distinct enough. Finally, some works evaluate ESS approaches by how well annotators are able to classify the synthesized emotions [49], [117], [122], [124], [125], [127].

As human evaluations are often costly and time-consuming, the community has attempted to supplement them with automatic ones. These evaluations are based on algorithmic measures that quantify different signal properties [169]. In the case of TTS, for example, BSD [171], PESQ [172], POLQA [173], or ITU-T Rec. P.563 [174] is often used to evaluate the quality of generated signals. No such standardized procedure exists yet for ESS, but several researchers are using distance metrics (e.g., Euclidean) between generated and target features, such as Mel spectra, or even using pretrained SER models to judge whether generated samples are correctly classified [175]. These metrics, though far from error-free, vastly speed up the development process of ESS approaches by providing quick feedback to researchers and are, thus, an integral part of the ESS ecosystem.

N. Datasets of Emotional Speech

In a data-driven paradigm, datasets become the cornerstone of successful models. A comprehensive overview of existing datasets of emotional speech used in ESS can be found in [17]. The authors mention five key desiderata for designing datasets that cover all conditions necessary for generalization: 1) increasing lexical variability, as emotional speech datasets are often recorded using a limited set of sentences; 2) introducing language variability, as ESS approaches might be expected to work for different languages and cultures; 3) promoting speaker variability, as acted datasets are typically recorded from a few actors and, thus, do not generalize well to new speakers; 4) controlling for confounders, such as different accents or demographics; and 5) regulating recording conditions, both to control unwanted confounders and to safeguard the quality of ground truth samples. However, this last factor can also act as an inhibiting factor for ESS applications that should generalize to different background environments; thus, we consider it a good restriction, while the field is still in its nascent stages, but one that must ultimately be abandoned as we transition to more realistic applications. The authors also introduce a new dataset, ESD, which is now being increasingly used by the community as a standard benchmark. Prior to the introduction of ESD, researchers used either small-scale datasets created explicitly for ESS [110], [113], [120] or relied on the standard SER datasets, such as IEMOCAP [176] and EMO-DB [177].

Given the small sizes of these datasets, some works have experimented with data augmentation to account for the scarcity of data [178]. Another viable way of dramatically increasing the size of available data is to utilize open resources. For example, with the advent of social media platforms, there exists now a vast quantity of readily available, high-quality, expressive speech data that could be tapped into for training future architectures. Furthermore, advances in automatic speech recognition (ASR) make it possible to obtain reliable transcriptions for those resources and use them as the input in an ESS pipeline. We expect these resources to play an increasingly important role in the synthesis of emotion and affect in general.

O. State-of-the-Art Performance

At the end of our overview of modern emotional speech synthesis approaches, one important question remains open: is ESS a solved problem? Recent works boasting average emotion similarity MOS scores of 4 for 6 [132], [134] or 7 [131] emotion categories certainly suggest that we are approaching a "WaveNet moment" for ESS as well, as the revolution started by van den Oord et al. [14] began with such MOS scores for naturalness. Accordingly, some works are showing subjective emotion recognition accuracies reaching up to 80% [49], [125]. Other works, however, feature much lower scores, dropping down to almost 50% recognition accuracy for four emotions [118], [122]. While one could easily dismiss the low-performing approaches as simply inferior, a closer look at the data used in each work reveals a more nuanced interpretation: Li et al. [132] and Lei et al. [134] used read speech datasets recorded by single, female authors specifically constructed for ESS, while Gao et al. [118] and Cao et al. [122] both used IEMOCAP [176], which includes improvized emotional speech.

This begs the question: how do we define success? This brings us back to the original question of what makes an affective agent. Success depends on the type of agent and the environment that they expect to operate. It depends on the number and kind of emotions that the agent is expected to support, the languages and cultures that it needs to cover, its malleability to user input, its robustness to different noise conditions, and so on. As ESS makes their journey out of research labs and into the real world, we expect fluctuations between periods of high performance on restricted conditions, followed by low valleys of MOS scores, as the application field is expanded, and evaluation criteria get increasingly stricter. Existing works show that the barrier of single-/few-speaker ESS systems with limited acted emotions on read speech has been breached, but we are only now approaching the frontier of naturalistic emotions, as most recent works are still relying on acted emotional data.

V. DISCUSSION

Our overview has shown that emotional speech synthesis is a rapidly growing field, which is being heavily influenced by the deep learning era of AI. In Section IV-O, we argued that, while DL constitutes an immense leap forward compared to previous approaches, ESS remains far from solved. In Section V, we highlight the main limitations, discuss whether the ESS problem should be solved at all given the ethical considerations that it raises, and, finally, outline some promising areas of future research.

A. Main Limitations

As seen in Section IV-M, ESS is still plagued by a lack of holistic, standardized evaluation protocols. In particular, there is a poignant lack of automatic evaluation benchmarks that allow a fair comparison of different approaches. As seen in other fields of AI, benchmarks become the driving force, which guides new advances. In contrast, even though significant progress has been made in recent years in ESS, this progress is hard to distill in a single leaderboard, which highlights the most promising future directions. More importantly, for any new algorithm that needs to be compared with the state of the art, researchers have to revert to costly human evaluations that hinder the rapid advance of the field. This makes it harder to iterate over new ideas and ascertain the impact of a proposed improvement. It is, however, a problem that can be easily solved by a focused effort of the community to create those standardized evaluation benchmarks.

A more serious challenge is achieving the amount of controllability required by downstream applications. Disentanglement of all confounding factors that influence a speech utterance remains the "holy grail" of emotional speech synthesis (and, for that matter, analysis too). Without proper disentanglement, ESS methods will be unable to yield a suitable set of "knobs" that an end-user can twist to generate the appropriate emotion. This problem also plagues state-of-the-art SER architectures [159], where models learn an entangled representation of linguistics and acoustics [179]. As ESS is scaled up to naturalistic datasets with a bigger lexical variability, we expect this issue to arise there as well.

Overrepresentation of a few "dominant" cultures and languages is another problem; while it is motivated by pragmatic reasons, namely, the availability of data; it, nevertheless, limits the applicability of the developed approaches. While research in related fields, such as ASR, shows that algorithms will generalize well to new languages once trained with data from those languages, it remains a challenge to procure data of such quantity for most of those. The use of more data-efficient methods to drastically cut down on the demand for data is still an open issue in the deep learning era of AI though we expect advances in neighboring fields to trickle down to ESS as well. However, this overall lack of cultural representation also raises ethical concerns as to whether ESS research can be universally applied and, thus, should be seriously considered by the community, besides the point of finding the data (see Section V-B).

Finally, we would be amiss not to point out the fact that contemporary affective computing research shies away from the problem of endowing machines with the capacity to have emotions. Thus, ESS approaches adhere to the "fake it until you make it" mantra, whereby HCI agents simulate the presence of emotions by appropriately modulating their voice. However, as research in human emotions has shown, there is a noticeable difference between acted and natural emotions (which can only, if ever, be circumvented by the best of actors) [180]. Therefore, it could be that the gap between humans and machines cannot be bridged until the latter also acquires the ability to simulate realistic emotions.

B. Ethical Considerations

In recent years, it has become increasingly evident that, just because artificial intelligence methods *can* do something, it does not necessarily mean that they *should* do it. This is also a central question in the field of emotional speech synthesis. While the potential to dramatically improve human–computer interaction, assist speaking-impaired individuals, and give voices to the intelligent agents of tomorrow is thrilling, there are several societal challenges facing our community in the here and now.

The most poignant of those issues is the rise of "deep fakes" (AI-fabricated videos of people saying or doing something that they have never said or done in real life) [181]. With the rapid advances in emotional speech synthesis, it is not far-fetched to assume that future "deep fakes" are not only going to change the linguistic content of targeted speakers but also their emotional one. This vastly increases the capabilities of malignant actors to spread disinformation about, or defame, a particular individual, even without changing their choice of words. For example, simply changing the tone of a politician who refers to a particular demographic group to sound sarcastic or derogatory could incur substantial damage to their public image.

A similar, more insidious approach would be to adapt the perceived personality of the target speaker. This can be used to make a particular candidate more or less appealing, or even to cast a whole demographic in a particular light, by manipulating the personalities of its spokespeople, e.g., to be seen as more aggressive or submissive. One particular example is that of voice assistants: as criticized in recent a UNESCO report [182], the initial design of several voice assistants was to show submissiveness, even in the face of blatant abuse, reinforcing notions of outdated "female servility." This case study shows how biases can be perpetuated through technological products in particular when those relate to a simulation of behavior and personality. This potential to transform public opinion through the use of targeted misinformation represents a major threat to societies around the world and would be vastly exacerbated by the improvement of ESS algorithms.

A final aspect of whether we "should" do ESS is whether we want conversational agents to be emotional. This will give them the unprecedented capability to influence our own emotions, perhaps in ways we would prefer to avoid. For example, agents whose purpose is to elicit more sales could adapt their voice to appear more trustworthy or friendly, thus subverting the buyer's will. Moreover, a related question is whether artificial agents should be clearly distinguishable from humans; the EU White Paper on Artificial Intelligence explicitly states that humans should be made cognizant of the fact that they are interacting with an artificial entity under all circumstances [183], but the question remains if that is sufficient to mitigate the potential dangers that could arise from "overhumanizing" those entities.

The question of "should" does, however, not cover the degree to which we "can." As is evident from the approaches presented here, full-blown conversational agents with the capacity to accurately and naturally convey emotion are increasingly on the way. Still, there are still a lot of critical considerations to be addressed. The first one is *generalizability*: do we cover all different cultures? and do we accurately represent all individuals? The second one is *privacy*: whence do we source our data from? The third one is *correctness*: is our evaluation sufficient?

None of these questions can be answered satisfactorily (yet). Research in ESS is being targeted to a small number of languages and cultures, the ones typically available in existing datasets, such as English or Chinese. Moreover, the emotions in these datasets and the corresponding synthesized samples are typically annotated by individuals of particular demographics (often students in the case of university research). This calls into question whether we are accurately capturing all the nuances of emotional speech across different cultures.

Emotions are also one of the most precious aspects of human experiences. Sourcing the vast quantities of those required by contemporary approaches is challenging without violating privacy. In particular, collecting negative emotions in realistic scenarios requires us to infringe on the most private moments of an individual, such as the heartbreaking loss of a loved one. Acted data can only get us part of the way there, but how we take the next step needed for naturalistic emotions remains an open, and challenging, question.

Evaluation is perhaps the easiest of the three questions. Decades of research on the perception of emotional speech provide a solid background from which to start. Co-opting those approaches for the evaluation of synthesized emotional speech and adopting best practices from the sister domain of TTS seem like a realistic goal.

Overall, it seems obvious that ESS leads to very serious ethical, legal, and social impact (ELSI) challenges. A full consideration of ELSI aspects cannot be given here, as it is too wide in scope for transformative technology, such as ESS. However, specifically for the field of computational paralinguistics, the reader is referred to [184].

C. Future Perspectives

After the tremendous advances that the TTS field saw in the last few years, ESS seems poised to become the next frontier for the speech synthesis community. Aside from tackling existing challenges and addressing the ethical considerations raised in Sections III and IV, we expect a few methodological advances to capture the interest of the community.

Synthesizing emotional vocal bursts is one of them. In the now famous promotional video for Google Assistant,[5] the crowd erupted in cheers as the assistant assured the hairdresser that "taking one second" to look for an appointment was fine with a mere "Mm-hmm." This illustrates how vocal bursts are essential components of emotional responses [53]. Synthesizing them was already the topic of the 2020 ExVo Challenge.[6] The best-performing approach, which used StyleGAN2, already achieved promising results that highlight the potential of this line of research [185].

Similarly, as stated in our introduction, conveying emotions is but one aspect of an affective agent. Endowing the agent with an artificial personality is another area, which has been pursued several decades [186]. This topic has been recently revived in the context of big language models, which can be adapted to emulate a specific personality [187]. As personality has been also shown to manifest in speech signals [188], it is an evident next step to introduce it to conversational agents as well [189]. In general, as exemplified by the tasks featured in the Computational Paralinguistics Challenge,[7] there exist a plethora of speaker states and traits, which can be modeled from the speech: deception, sincerity, nativeness, cognitive load, likability, interest, and others are all variables that could be added to the capabilities of affective agents.

Personalization is expected to be another major aspect of future ESS systems. Both the expression [190], [191], [192] and the perception [193] of emotion show individualistic effects, which are currently underexploited in the ESS field. Future approaches can benefit a lot from adopting a similar mindset and adapt the production of emotional speech to a style that fits both the speaker and the listener. Such an interpersonal adaptation effect is also seen in human conversations and is a necessary step to foster communication [194].

Finally, as future affective agents find their way out of their academic research sandboxes and into the real world, they will be forced to interact with other entities—artificial and human alike. This will form a natural breeding ground for interactions, which can be accordingly classified as "successful" or not, depending on the goals of the agent. Coupled with effective SER capabilities, these interactions constitute a natural *reward signal*, which can be further utilized by their agent to improve their ESS and SER capacities in a lifelong reinforcement learning setup, which still remains an elusive goal for the field of affective computing [195]. An overture to this exciting domain can already be found in intelligent dialog generation, where reinforcement learning is already being used to adjust the linguistic style of an agent [196] or to learn backchanneling responses [197], [198]. We expect this paradigm to be more widely used in ESS in the near future.

VI. CONCLUSION

We have presented an overview of recent advances in the synthesis of affective speech, including affective voice conversion. Deep learning is paving the way for considerable advances in this field and laying the foundation for the affective conversational agents of tomorrow. Most work has focused on categorical emotions, using, in particular, acted datasets of read speech. The community has mostly concentrated on modifying acoustic features, a form of affective voice conversion, but there is recently a renaissance of ESS approaches that directly map text to acoustics. Accordingly, we are seeing an increasing consolidation of advances in TTS and a move toward more "E2E" affective synthesis. Finally, following recent successes in the conversion of one emotion category to another, several works are now focusing on a more granular control of the intensity, thus increasing the controllability of EVC methods.

As main challenges to existing approaches, we have identified the absence of naturalistic emotions in the most widely used corpora used for emotional speech synthesis, the overrepresentation of a few cultures and languages in

[5] https://www.youtube.com/watch?v=yDI5oVn0RgM

[6] https://www.competitions.hume.ai/exvo2022

[7] www.compare.openaudio.eu

emotional datasets, the issue of disentangling the different latent factors that influence speech, and the inherent limitations of an approach that tries to imitate, rather than simulate emotions. Another major challenge is the adherence to ethical rules, as machines that can simulate affect in all their manifestations, such as emotion and personality, can pose serious threats to societies in the era of "fake news." Nevertheless, we believe that concentrated efforts by the community can overcome these barriers and help realize the full potential of affective agents.

REFERENCES

[1] A. J. Fiannaca, A. Paradiso, J. Campbell, and M. R. Morris, "Voicesetting: Voice authoring UIs for improved expressivity in augmentative communication," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Montreal, QC, Canada, Apr. 2018, pp. 1–12.

[2] M. Minsky, *Society of Mind*. New York, NY, USA: Simon and Schuster, 1988.

[3] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[4] A. Batliner and B. Möbius, "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?" in *The Integration of Phonetic Knowledge in Speech Technology*. Springer, 2005, pp. 21–44.

[5] Y. Stylianou, "Voice transformation: A survey," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taiwan, Apr. 2009, pp. 3585–3588.

[6] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.

[7] I. R. Murray, "Simulating emotion in synthetic speech," Ph.D. dissertation, Dept. Math. Comput. Sci., Univ. Dundee, Dundee, U.K., 1989.

[8] J. E. Cahn, "Generating expression in synthesized speech," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1989.

[9] J. E. Cahn, "The generation of affect in synthesized speech," *J. Amer. Voice I/O Soc.*, vol. 8, no. 1, p. 1, 1990.

[10] Y. Kitahara and Y. Tohkura, "Prosodic control to express emotions for man-machine speech interaction," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 75, no. 2, pp. 155–163, 1992.

[11] B. Granström, "The use of speech synthesis in exploring different speaking styles," *Speech Commun.*, vol. 11, nos. 4–5, pp. 347–355, Oct. 1992.

[12] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, Feb. 1993.

[13] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[14] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*. Sunnyvale, CA, USA: ISCA, 2016, p. 125.

[15] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021, *arXiv:2106.15561*.

[16] M. Schröder, "Emotional speech synthesis: A review," in *Proc. 7th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2001, pp. 561–564.

[17] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, Feb. 2022.

[18] Z. Yang, X. Jing, A. Triantafyllopoulos, M. Song, I. Aslan, and B. W. Schuller, "An overview & analysis of sequence-to-sequence emotional voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Incheon, South Korea: ISCA, Sep. 2022, pp. 1–5.

[19] R. Li, Z. Wu, Y. Huang, J. Jia, H. Meng, and L. Cai, "Emphatic speech generation with conditioned input layer and bidirectional LSTMS for expressive speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Alberta, Apr. 2018, pp. 5129–5133.

[20] L. Liu et al., "Controllable emphatic speech synthesis based on forward attention for expressive speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 410–414.

[21] S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Emotional voice conversion using a hybrid framework with speaker-adaptive DNN and particle-swarm-optimized neural network," *IEEE Access*, vol. 8, pp. 74627–74647, 2020.

[22] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affective Comput.*, vol. 5, no. 2, pp. 101–111, Apr./Jun. 2014.

[23] M. Schröder et al., "Building autonomous sensitive artificial listeners," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 165–183, Apr./Jun. 2012.

[24] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6593–6597.

[25] K. Lakhotia et al., "On generative spoken language modeling from raw audio," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1336–1354, Dec. 2021.

[26] A. Polyak et al., "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Brno, Czech Republic: ISCA, Aug. 2021, pp. 3616–3619.

[27] E. Kharitonov et al., "Text-free prosody-aware generative spoken language modeling," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: ACL, 2022, pp. 8666–8681.

[28] J. Monge Alvarez et al., "CAMNet: A controllable acoustic model for efficient, expressive, high-quality text-to-speech," *Appl. Acoust.*, vol. 186, Jan. 2022, Art. no. 108439.

[29] J. P. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, *Progress in Speech Synthesis*. Springer, 1997.

[30] N. Campbell, "Expressive/affective speech synthesis," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 505–518.

[31] K. R. Scherer, "Appraisal theory," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power, Eds. Hoboken, NJ, USA: Wiley, 1999, pp. 637–663.

[32] P. C. Ellsworth and K. R. Scherer, *Appraisal Processes in Emotion*. London, U.K.: Oxford Univ. Press, 2003.

[33] R. Lan et al., "Chinese emotional dialogue response generation via reinforcement learning," *ACM Trans. Internet Technol.*, vol. 21, no. 4, pp. 1–17, Nov. 2021.

[34] Y. Chiba, T. Nose, T. Kase, M. Yamanaka, and A. Ito, "An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system," in *Proc. 19th Annu. SIGdial Meeting Discourse Dialogue*. Melbourne, VIC, Australia: ACL, 2018, pp. 371–375.

[35] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1. New Orleans, LA, USA: AAAI, 2018, pp. 1-8.

[36] Z. Bucinca, Y. Yemez, E. Erzin, and M. Sezgin, "AffectON: Incorporating affect into dialog generation," *IEEE Trans. Affect. Comput.*, early access, Dec. 8, 2020, doi: 10.1109/TAFFC.2020.3043067.

[37] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective," *J. Exp. Psychol., Gen.*, vol. 113, no. 3, pp. 464–486, Sep. 1984.

[38] K. R. Scherer, "Towards a prediction and data driven computational process model of emotion," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 279–292, Apr. 2021.

[39] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.

[40] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.

[41] M. Schroder, "Expressing degree of activation in synthetic speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1128–1136, Jul. 2006.

[42] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Commun.*, vol. 102, pp. 54–67, Sep. 2018.

[43] W. Zhang, H. Yang, and P. Zhi, "Emotional speech synthesis based on DNN and PAD emotional state model," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Taiwan, Nov. 2018, pp. 41–45.

[44] K. R. Scherer et al., "Psychological models of emotion," *Neuropsychol. Emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[45] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annu. Rev. Psychol.*, vol. 41, no. 1, pp. 417–440, 1990.

[46] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Social Psychol.*, vol. 70, no. 3, p. 614, 1996.

[47] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," in *Handbook of Emotions*, vol. 2. 2000, pp. 220–235.

[48] N. Campbell and P. Mokhtari, "Voice quality: The 4th prosodic dimension," in *Proc. Int. Congr. Phonetic Sci.* Barcelona, Spain: International Phonetic Association, 2003, pp. 2417–2420.

[49] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42674–42687, 2021.

[50] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Trans. Affect. Comput.*, early access, May 19, 2022, doi: 10.1109/TAFFC.2022.3175578.

[51] M. Tahon, G. Lecorvé, and D. Lolive, "Can we generate emotional pronunciations for expressive speech synthesis?" *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 684–695, Oct. 2020.

[52] A. Baird et al., "The ICML 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts," 2022, *arXiv:2205.01780*.

[53] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization," *Amer. Psychol.*, vol. 74, no. 6, p. 698, 2019.

[54] H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgang von Kempelen," *J. Acoust. Soc. Amer.*, vol. 22, no. 2, pp. 151–166, 1950.

[55] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, no. 4, pp. 452–460, Apr. 1976.

[56] J. Allen, S. Hunnicutt, R. Carlson, and B. Granstrom, "MITalk-79: The 1979 MIT text-to-speech system," *J. Acoust. Soc. Amer.*, vol. 65, no. S1, p. S130, Jun. 1979.

[57] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, 1987.

[58] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using

diphones," *Speech Commun.*, vol. 9, nos. 5–6, pp. 453–467, Dec. 1990.

[59] R. A. Khan and J. S. Chitode, "Concatenative speech synthesis: A review," *Int. J. Comput. Appl.*, vol. 136, no. 3, pp. 1–6, Feb. 2016.

[60] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3. Istanbul, Turkey, Jun. 2000, pp. 1315–1318.

[61] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[62] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[63] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[64] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in Psychology*. Amsterdam, The Netherlands: Elsevier, 1997, vol. 121, pp. 471–495.

[65] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[66] S. O. Arik et al., "Deep voice: Real-time neural text-to-speech," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, D. Precup and Y. W. Teh, Eds, Sydney, NSW, Australia, 2017, pp. 195–204.

[67] A. Gibiansky et al., "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[68] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 4006–4010.

[69] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4779–4783.

[70] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, 2020, pp. 7586–7598.

[71] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.* Honolulu, HI, USA: AAAI, 2019, pp. 6706–6713.

[72] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[73] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, 2020, pp. 1–15.

[74] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Scottsdale, AZ, USA, 2013, pp. 1–14.

[75] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[76] A. Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 3918–3926.

[77] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[78] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[79] K. Kumar et al., "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 14881–14892.

[80] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6199–6203.

[81] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17022–17033.

[82] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*. Vancouver, BC, Canada, 2018, pp. 1–16.

[83] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3617–3621.

[84] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, 2021, pp. 8599–8608.

[85] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[86] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 132–157, 2021.

[87] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[88] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, no. 2, pp. 147–158, 1989.

[89] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[90] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, B. Yegnanarayana, Ed. Hyberabad, India: ISCA, Sep. 2018, pp. 1978–1982.

[91] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process.*, Toulouse, France, May 2006, pp. 77–80.

[92] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5274–5278.

[93] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Jeju, South Korea, Dec. 2016, pp. 1–6.

[94] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Stuttgart, Germany, Dec. 2018, pp. 266–273.

[95] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2100–2104.

[96] S. Liu et al., "Transferring source style in non-parallel voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, H. Meng, B. Xu, and T. F. Zheng, Eds. Shanghai, China: ISCA, Oct. 2020, pp. 4721–4725.

[97] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6805–6809.

[98] Y. Li, K. A. Lee, Y. Yuan, H. Li, and Z. Yang, "Many-to-Many voice conversion based on bottleneck features with variational autoencoder for non-parallel training data," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Honolulu, HI, USA, Nov. 2018, pp. 829–833.

[99] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251.

[100] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6820–6824.

[101] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8789–8797.

[102] A. Moritani, S. Sakamoto, R. Ozaki, H. Kameokal, and T. Taniguchi, "StarGAN-based emotional voice conversion for Japanese phrases," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2021, pp. 836–840.

[103] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*, Beijing, China, 2000, pp. 1–6.

[104] A. W. Black, "Unit selection and emotional speech," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Geneva, Switzerland, Sep. 2003, pp. 1649–1652.

[105] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, nos. 1–2, pp. 161–187, Apr. 2003.

[106] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *Proc. Int. Conf. Speech Prosody*. Nara, Japan: ISCA, 2004, pp. 413–416.

[107] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[108] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, N. Morgan, Ed. San Francisco, CA, USA: ISCA, Sep. 2016, pp. 2453–2457.

[109] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," 2017, *arXiv:1711.05447*.

[110] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, May 2018.

[111] X. An, Y. Wang, S. Yang, Z. Ma, and L. Xie, "Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 184–191.

[112] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6950–6954.

[113] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis based on style embedded Tacotron2 framework," in *Proc. 34th Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC)*, Jeju, South Korea, Jun. 2019,

pp. 1–4.

[114] R. Shankar, J. Sager, and A. Venkataraman, “A multi-speaker emotion morphing model using highway networks and maximum likelihood objective,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Graz, Austria: ISCA, Sep. 2019, pp. 2848–2852.

[115] F. Bao, M. Neumann, and N. T. Vu, “CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, G. Kubin and Z. Kacic, Eds. Graz, Austria: ISCA, Sep. 2019, pp. 2828–2832.

[116] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, “Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform F0 features,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1535–1548, Oct. 2019.

[117] C. Robinson, N. Obin, and A. Roebel, “Sequence-to-sequence modelling of F0 for speech emotion conversion,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, U.K., May 2019, pp. 6830–6834.

[118] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, “Nonparallel emotional speech conversion,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, G. Kubin and Z. Kacic, Eds. Graz, Austria: ISCA, Sep. 2019, pp. 2858–2862.

[119] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, “Using VAEs and normalizing flows for one-shot text-to-speech synthesis of expressive speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6179–6183.

[120] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, “Emotional voice conversion using multitask learning with text-to-speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7774–7778.

[121] G. Rizos, A. Baird, M. Elliott, and B. Schuller, “StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 3502–3506.

[122] Y. Cao, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, “Nonparallel emotional speech conversion using VAE-GAN,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Shanghai, China: ISCA, Oct. 2020, pp. 3406–3410.

[123] Y. Lei, S. Yang, and L. Xie, “Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis,” in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 423–430.

[124] B. Schnell and P. N. Garner, “Improving emotional TTS with an emotion intensity input from unsupervised extraction,” in *Proc. 11th ISCA Speech Synth. Workshop (SSW)*. Brno, Czech Republic: ISCA, Aug. 2021, pp. 60–65.

[125] R. Liu, B. Sisman, and H. Li, “Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability,” 2021, *arXiv:2104.01408*.

[126] Z. Du, B. Sisman, K. Zhou, and H. Li, “Expressive voice conversion: A joint framework for speaker identity and emotional style transfer,” in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Cartagena, Colombia, Dec. 2021, pp. 594–601.

[127] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, “Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 5734–5738.

[128] X. Wu et al., “Exemplar-based emotive speech synthesis,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 874–886, 2021.

[129] F. Kreuk et al., “Textless speech emotion conversion using discrete and decomposed representations,” 2021, *arXiv:2111.07402*.

[130] X. Chen, X. Xu, J. Chen, Z. Zhang, T. Takiguchi, and E. R. Hancock, “Speaker-independent emotional voice conversion via disentangled representations,” *IEEE Trans. Multimedia*, early access, Nov. 16, 2022, doi: 10.1109/TMM.2022.3222646.

[131] G. Zhang et al., “iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre,” 2022, *arXiv:2206.14866*.

[132] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, “Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1448–1460, 2022.

[133] R. Liu, B. Sisman, G. Gao, and H. Li, “Expressive TTS training with frame and style reconstruction loss,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1806–1818, 2021.

[134] Y. Lei, S. Yang, X. Wang, and L. Xie, “MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 853–864, 2022.

[135] Y. Feng, P. Duan, Y. Zi, Y. Chen, and S. Xiong, “Fusing acoustic and text emotional features for expressive speech synthesis,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Taiwan, Jul. 2022, pp. 1–6.

[136] S. An, Z. Ling, and L. Dai, “Emotional statistical parametric speech synthesis using LSTM-RNNs,” in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Kuala Lumpur, Malaysia, Dec. 2017, pp. 1613–1616.

[137] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proc. Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 933–941.

[138] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017, pp. 1–14.

[139] K. Zhou, B. Sisman, and H. Li, “Transforming spectrum and prosody for emotional voice conversion with non-parallel training data,” 2020, *arXiv:2002.00198*.

[140] S. Liu, Y. Cao, and H. Meng, “Emotional voice conversion with cycle-consistent adversarial network,” 2020, *arXiv:2004.03781*.

[141] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, “An improved CycleGAN-based emotional voice conversion model by augmenting temporal dependency with a transformer,” *Speech Commun.*, vol. 144, pp. 110–121, Oct. 2022.

[142] X. He, J. Chen, G. Rizos, and B. W. Schuller, “An improved StarGAN for emotional voice conversion: Enhancing voice quality and data augmentation,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Brno, Czech Republic: ISCA, Aug. 2021, pp. 821–825.

[143] K. Zhou, B. Sisman, R. Liu, and H. Li, “Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 920–924.

[144] Y. Lei, S. Yang, X. Zhu, L. Xie, and D. Su, “Cross-speaker emotion transfer through information perturbation in emotional speech synthesis,” *IEEE Signal Process. Lett.*, vol. 29, pp. 1948–1952, 2022.

[145] G. Zhang, S. Qiu, Y. Qin, and T. Lee, “Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis,” in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Jan. 2021, pp. 1–5.

[146] D. R. Ladd, K. E. A. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, “Evidence for the independent function of intonation contour type, voice quality, and $F0$ range in signaling speaker affect,” *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 435–444, Aug. 1985.

[147] K. R. Scherer, “Personality inference from voice quality: The loud voice of extroversion,” *Eur. J. Social Psychol.*, vol. 8, no. 4, pp. 467–487, Oct. 1978.

[148] F. Eyben et al., “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2015.

[149] A. Huang, F. Bao, G. Gao, Y. Shan, and R. Liu, “Mongolian emotional speech synthesis based on transfer learning and emotional embedding,” in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Singapore, Dec. 2021, pp. 78–83.

[150] R. Skerry-Ryan et al., “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 4693–4702.

[151] Y. Wang et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 5180–5189.

[152] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 623–627.

[153] K. He, C. Sun, R. Zhu, and L. Zhao, “Multi-speaker emotional speech synthesis with limited datasets: Two-stage non-parallel training strategy,” in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Beijing, China, Apr. 2022, pp. 545–548.

[154] S. Moon, S. Kim, and Y.-H. Choi, “MIST-tacotron: End-to-end emotional speech synthesis using mel-spectrogram image style transfer,” *IEEE Access*, vol. 10, pp. 25455–25463, 2022.

[155] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, “Emotional speech synthesis with rich and granularized control,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7254–7258.

[156] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, “An effective style token weight control technique for end-to-end emotional speech synthesis,” *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, Sep. 2019.

[157] M. Osman, “EVASS: Emotional variational end-to-end speech synthesis with semi-supervised and adverserial learning,” in *Proc. 2nd Int. Mobile, Intell., Ubiquitous Comput. Conf. (MIUCC)*, Cairo, Egypt, May 2022, pp. 97–103.

[158] D. Stanton, Y. Wang, and R. Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 595–602.

[159] J. Wagner et al., “Dawn of the transformer era in speech emotion recognition: Closing the valence gap,” 2022, *arXiv:2203.07378*.

[160] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions,” in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, San Antonio, TX, USA, Oct. 2017, pp. 248–255.

[161] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[162] C. Lu, X. Wen, R. Liu, and X. Chen, “Multi-speaker emotional speech synthesis with fine-grained prosody modeling,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 5729–5733.

[163] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: The Munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM Int. Conf. Multimedia*, Florence, Italy, Oct. 2010, pp. 1459–1462.

[164] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, “Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2017, no. 1, p. 18, Dec. 2017, doi: 10.1186/s13636-017-0116-2.

[165] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, “Neutral-to-emotional voice conversion with cross-wavelet transform f0 using generative

adversarial networks," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, 2019, Art. no. e10.

[166] M. Elgaar, J. Park, and S. W. Lee, "Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7769–7773.

[167] F. Yang, J. Luan, and Y. Wang, "Improving emotional speech synthesis by using SUS-constrained VAE and text encoder aggregation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 8302–8306.

[168] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, Mar. 2016.

[169] F. Hinterleitner, "Auditory and instrumental quality evaluation metrics," in *Quality of Synthetic Speech*. Springer, 2017, pp. 19–36.

[170] Y.-T. Huang and H. T. Lawless, "Sensitivity of the ABX discrimination test," *J. Sensory Stud.*, vol. 13, no. 2, pp. 229–239, Jul. 1998.

[171] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[172] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[173] J. G. Beerends et al., "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement. Part I—Temporal alignment," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[174] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[175] A. Baird et al., "A prototypical network approach for evaluating generated emotional speech," in *Proc. Interspeech*. Brno, Czech Republic: ISCA, Aug. 2021, pp. 3161–3165.

[176] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[177] F. Burkhardt et al., "A database of German emotional speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 5. Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[178] M. Lajszczak et al., "Distribution augmentation for low-resource expressive text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8307–8311.

[179] A. Triantafyllopoulos et al., "Probing speech emotion recognition transformers for linguistic knowledge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2022, pp. 146–150.

[180] R. Cowie, E. Douglas-Cowie, M. McRorie, I. Sneddon, L. Devillers, and N. Amir, "Issues in data collection," in *Emotion-Oriented Systems*, 2011, pp. 197–212.

[181] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Rev.*, vol. 107, p. 1753, 2019.

[182] M. West, R. Kraut, and H. E. Chew, "I'd blush if I could: Closing gender divides in digital skills through education," Ministry Educ., Peru, Tech. Rep., 2019. [Online]. Available: http://repositorio.minedu.gob.pe/handle/20.500.12799/6598

[183] "White paper on artificial intelligence: A European approach to excellence and trust," EU Commission, Brussels, Belgium, Tech. Rep. 65, 2020.

[184] A. Batliner, S. Hantke, and B. Schuller, "Ethics and good practice in computational paralinguistics," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1236–1253, Jul./Sep. 2022.

[185] M. Jiralerspong and G. Gidel, "Generating diverse vocal bursts with StyleGAN2 and MEL-spectrograms," 2022, *arXiv:2206.12563*.

[186] B. L. Brown, W. J. Strong, and A. C. Rencher, "Perceptions of personality from speech: Effects of manipulations of acoustical parameters," *J. Acoust. Soc. Amer.*, vol. 54, no. 1, pp. 29–35, Jul. 1973.

[187] S. S. Keh and I.-T. Cheng, "Myers–Briggs personality classification and personality-specific language generation using pre-trained language models," 2019, *arXiv:1907.06333*.

[188] B. Schuller et al., "The INTERSPEECH 2012 speaker trait challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Portland, OR, USA: ISCA, Sep. 2012, pp. 1–4.

[189] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," in *Proc. Int. Workshop Affect. Interact.* Springer, 1999, pp. 150–165.

[190] H. Ritschel, I. Aslan, S. Mertes, A. Seiderer, and E. André, "Personalized synthesis of intentional and emotional non-verbal sounds for social robots," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Cambridge, U.K., Sep. 2019, pp. 1–7.

[191] A. Baird, S. Amiriparian, and B. Schuller, "Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*. Kuala Lumpur, Malaysia: IEEE, Sep. 2019, pp. 1–5.

[192] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shenzhen, China, Jul. 2021, pp. 1–6.

[193] A. Ando et al., "Speech emotion recognition based on listener adaptive models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ONT, Canada, Jun. 2021, pp. 6274–6278.

[194] S. Amiriparian et al., "Synchronization in interpersonal speech," *Frontiers Robot. AI*, vol. 6, p. 116, Nov. 2019.

[195] D. M. Schuller and B. W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Rev.*, vol. 13, no. 1, pp. 44–50, Jan. 2021.

[196] H. Ritschel, T. Baur, and E. Andr'e, "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Lisbon, Portugal, Aug. 2017, pp. 378–384.

[197] Ö. Z. Bayramoğlu, E. Erzin, T. M. Sezgin, and Y. Yemez, "Engagement rewarded actor-critic with conservative Q-learning for speech-driven laughter backchannel generation," in *Proc. Int. Conf. Multimodal Interact.*, Montreal, QC, Canada, Oct. 2021, pp. 613–618.

[198] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Training socially engaging robots: Modeling backchannel behaviors with batch reinforcement learning," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4 pp. 1840–1853, Oct./Dec. 2022.

ABOUT THE AUTHORS

Andreas Triantafyllopoulos received the Diploma degree in electrical and computer engineering (ECE), University of Patras, Patras, Greece, in 2017. He is currently working toward the Ph.D. degree at the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany.

He is currently a Research Assistant Researcher with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg. His current research focus is on deep learning methods for auditory intelligence and affective computing.

Björn W. Schuller (Fellow, IEEE) received the Diploma, Doctoral, and Habilitation degrees in electrical engineering (EE) and information technology (IT) from the Technical University of Munich (TUM), Munich, Germany, in 1999, 2006, and 2012, respectively.

He was an Adjunct Teaching Professor of EE/IT, TUM. He is currently a Full Professor of artificial intelligence and the Head of the Group on Language, Audio, & Music (GLAM), Imperial College London, London, U.K.; a Full Professor and the Chair of embedded intelligence for health care and well-being with the University of Augsburg, Augsburg, Germany; and the Co-Founding CEO and the Current CSO of audEERING, Gilching, Germany. He has (co)authored more than 1200 publications (more than 45k citations and H-index = 100+).

Dr. Schuller is a Fellow of the IEEE and Golden Core Awardee of the IEEE Computer Society, the British Computer Society (BCS), and the International Speech Communication Association (ISCA), the President-Emeritus of the Association for the Advancement of Affective Computing (AAAC), and a Senior Member of the Association for Computing Machinery (ACM). He was the Editor-in-Chief of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING among manifold further commitments and service to the community. He is also a Field Chief Editor of *Frontiers in Digital Health*.

Gökçe İymen received the B.S. degree in industrial engineering from Middle East Technical University, Ankara, Turkey, in 2019. She is currently working toward the M.S. degree in data science at Koç University, Istanbul, Turkey, where her research focuses on applications of deep learning for audio generation, especially adding affect to speech.

Metin Sezgin (Member, IEEE) graduated *summa cum laude* with Honors from Syracuse University, Syracuse, NY, USA, in 1999. He received the M.S. and Ph.D. degrees from the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2001 and 2006, respectively. He has held visiting posts at Harvard University, Cambridge, MA, USA, and Yale University, New Haven, CT, USA. He is currently an Associate Professor with the College of Engineering, Koç University, Istanbul, Turkey. His research has been supported by international and national grants, including grants from the European Research Council and Turk Telekom. His research interests include intelligent human–computer interfaces, multimodal sensor fusion, and human–computer interaction (HCI) applications of machine learning.

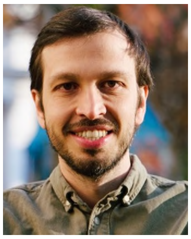
Dr. Sezgin was a recipient of the Career Award of the Scientific and Technological Research Council of Turkey.

Xiangheng He received the master's degree from Southeast University, Nanjing, China, in 2020. She is currently working toward the Ph.D. degree at the Group on Language, Audio, & Music (GLAM), Imperial College London, London, U.K.

She is currently a Research Assistant with the University of Augsburg, Augsburg, Germany. Her research focuses on affective computing and voice conversion.

Zijiang Yang (Student Member, IEEE) received the master's degree in information technology from the University of York, York, U.K., in 2016. He is currently working toward the Ph.D. degree at the University of Augsburg, Augsburg, Germany.

He is currently a Research Assistant with the University of Augsburg. His research focuses on deep learning, affective computing, and speech synthesis.



Panagiotis Tzirakis (Member, IEEE) received the Ph.D. degree from the Intelligent Behaviour Understanding Group (iBUG), Imperial College London, London, U.K., in 2021, where he focused on multimodal emotion recognition efforts.

He has published in top outlets, including *Information Fusion*, *International Journal of Computer Vision*, and several IEEE conference proceedings on topics including 3-D facial motion synthesis, multichannel speech enhancement, the detection of Gibbon calls, and emotion recognition from audio and video.

Shuo Liu received the M.Sc. degree in electrical engineering and information technology from the Technical University of Darmstadt (TUD), Darmstadt, Germany, in 2017. He is currently working toward the Ph.D. degree at the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany.

His current research interests include deep learning and machine learning algorithms for speech and audio processing, affective computing, and health-related applications.

Silvan Mertes received the M.Sc. degree in computer science from the Chair of Human-Centered Artificial Intelligence, University of Augsburg, Augsburg, Germany, in 2019, where he is currently working toward the Ph.D. degree.

His research focuses on generative adversarial learning for audio and image synthesis. Specifically, he explores how adversarial learning approaches can enhance datasets and explainability in different deep learning tasks.

Elisabeth André (Senior Member, IEEE) is currently a Full Professor of computer science and the Founding Chair of human-centered artificial intelligence with the University of Augsburg, Augsburg, Germany. She has a long track record in multimodal human–machine interaction, embodied conversational agents, social robotics, affective computing, and social signal processing.

Dr. André is a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina, and the CHI Academy. Most recently, she was named one of the ten most influential figures in the history of AI in Germany by the National Society for Informatics (GI). Her work has won many awards, including the Gottfried Wilhelm Leibniz Prize and the Most Important Research Funding Award in Germany. In 2013, she was awarded an EurAI (European Association for Artificial Intelligence) Fellowship. Since 2019, she has been serving as the Editor-in-Chief of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. She is a Co-Speaker of the Bavarian Research Association ForDigitHealth.

Ruibo Fu (Member, IEEE) received the B.E. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2015, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2020.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He has published more than ten papers in international conferences and journals, such as IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and Annual Conference of the International Speech Communication Association (INTERSPEECH). His research interests lie in speech synthesis and transfer learning.

Dr. Fu has won the Best Paper Award twice in National Conference on Man-Machine Speech Communication (NCMMSC) 2017 and 2019. He won the First Prize in the Personalized Speech Synthesis Competition held by the Ministry of Industry and Information Technology twice in 2019 and 2020. He also won the First Prize in the ICASSP 2021 Multi-Speaker Multi-Style Voice Cloning Challenge (M2VoC) Challenge.

Jianhua Tao (Senior Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001.

He is currently a Professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing. He has published more than 80 papers in major journals and proceedings, including IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech synthesis and coding methods, human–computer interaction, multimedia information processing, and pattern recognition.

Dr. Tao also serves as the Steering Committee Member of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. He received several awards from the fields' important conferences, such as Eurospeech or National Conference on Man-Machine Speech Communication (NCMMSC). He serves as the Chair or a Program Committee Member of several major conferences, including International Conference on Pattern Recognition (ICPR), International Conference on Affective Computing and Intelligent Interaction (ACII), International Conference on Multimodal Interaction (ICMI), International Symposium Chinese Spoken Language Processing (ISCSLP), or NCMMSC. He also serves as Associate Editor for *Journal on Multimodal User Interface* and *International Journal of Synthetic Emotions* and the Deputy Editor-in-Chief for *Chinese Journal of Phonetics*.