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

Andreas Triantafyllopoulos, *Student Member, IEEE*, 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*, 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. But 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 synthesised 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 which underlies most of the recent advances in artificial intelligence, is spearheading these efforts. In the present overview, we outline ongoing trends and summarise state-of-the-art approaches in an attempt to provide a comprehensive overview of this exciting field.**

***Index Terms*—Affective Computing, Speech Synthesis, Emotional Voice Conversion, Deep Learning.**

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 to 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, their emotions, their unique worldview. People who suffer from similar conditions understand that the mere verbalisation of their words using assistive technologies is not enough to give them back their voice. 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 to 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, Marvin Minsky [2]: "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions." And any being that has emotions requires an avenue to express them.

Affective computing is the subfield of AI that concerns itself with the computational modelling, understanding, and expression of emotions [3]. One of its primary goals is to facilitate more natural human-computer interaction (HCI) through the modelling of affect, which is a key component of human behaviour. 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 which attempts to generate realistic-sounding affective speech. We define affective speech synthesis as a sub-field 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 remains the primary pursuit of current research efforts, we will concentrate our review on this aspect of affective synthesis. We also consider both the general case of synthesising 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 synthesised 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 have primarily focused on rule-based

A. Triantafyllopoulos, B. Schuller, X. He, Z. Yang, S. Liu are with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.

B. Schuller, P. Tzirakis are with GLAM – the Group on Language, Audio, & Music, Imperial College London, SW7 2AZ London, UK.

G. İymen, Metin Sezgin are with the KUIS AI Lab., College of Engineering, Koç University, Istanbul, Turkey.

S. Mertes, E. André are the Chair of Human-Centered Artificial Intelligence, University of Augsburg, Germany.

R. Fu, J. Tao are with the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Manuscript received September 30, 2022

[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.

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 Oord *et al.* [14] revolutionised the field of text-to-speech synthesis (TTS), signalling the advent of the deep learning (DL) era and, more generally, solidifying the switch to a data-driven paradigm, where a mapping from text to speech is *learnt* using data. Similar approaches are now spearheading research in affective speech synthesis as well [13].

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 summarise recent efforts, synthesise 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 focus on 'mere' TTS [15], and older affective speech synthesis reviews which have become largely obsolete in the deep learning era [16], or newer ones which are more limited in scope [17, 18].

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. We then give a brief introduction on deep speech synthesis in general, followed by a thorough review of deep emotional speech synthesis and conversion. Finally, we summarise major observations and outline potential avenues for future research.

II. Affective speech generation

In this section, 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 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 on 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 emotional voice conversion 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 modelling of affect. Finally, we introduce the acoustic correlates of affect which have guided the 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 which more narrowly define affect as the manifestation of a subjectively experienced emotion [19]. Thus, in our review, affective speech is speech which encapsulates all possible paralinguistic traits and states [6].

Seen from that perspective, the majority of research in affective speech synthesis has been actually devoted to ESS, with considerably less emphasis on personality and other states or traits [6, 20]. This discrepancy is even more pronounced in the ongoing deep learning era, with far more work devoted to ESS than any other aspects of affect. For this reason, we focus our review on emotional speech synthesis and leave a discussion of other aspects of affect for Section V-C.

We note that several recent works investigate 'expressive' TTS [21, 22, 23, 24, 25]. This term is increasingly appearing on TTS papers to mean the control of prosody and other factors not related to content. However, it is often the case that this control is applied manually to change the style of the synthesised utterance. It is thus missing the explicit link to emotion that we consider critical for ESS. Naturally, prosody is highly related to emotion; therefore, any manipulation of it might result in a change in the perceived emotion. Yet, expressive TTS methods tend to leave out any evaluation of the emotional content of the synthesised utterances. 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 [8]. Indeed, the mechanisms used to induce expressivity are almost identical to those used to induce emotionality, albeit the latter explicitly use emotional information for guidance. We consider this a necessary prerequisite for successful ESS.

B. Affective agent model

While the majority of affective speech generation works are concerned with the task of endowing a synthesised utterance with appropriate emotional inflections, this is but the last step in the pipeline of an affective agent[3]. Figure 1 shows a coarse model of what is at play in an affective computing application.

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

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

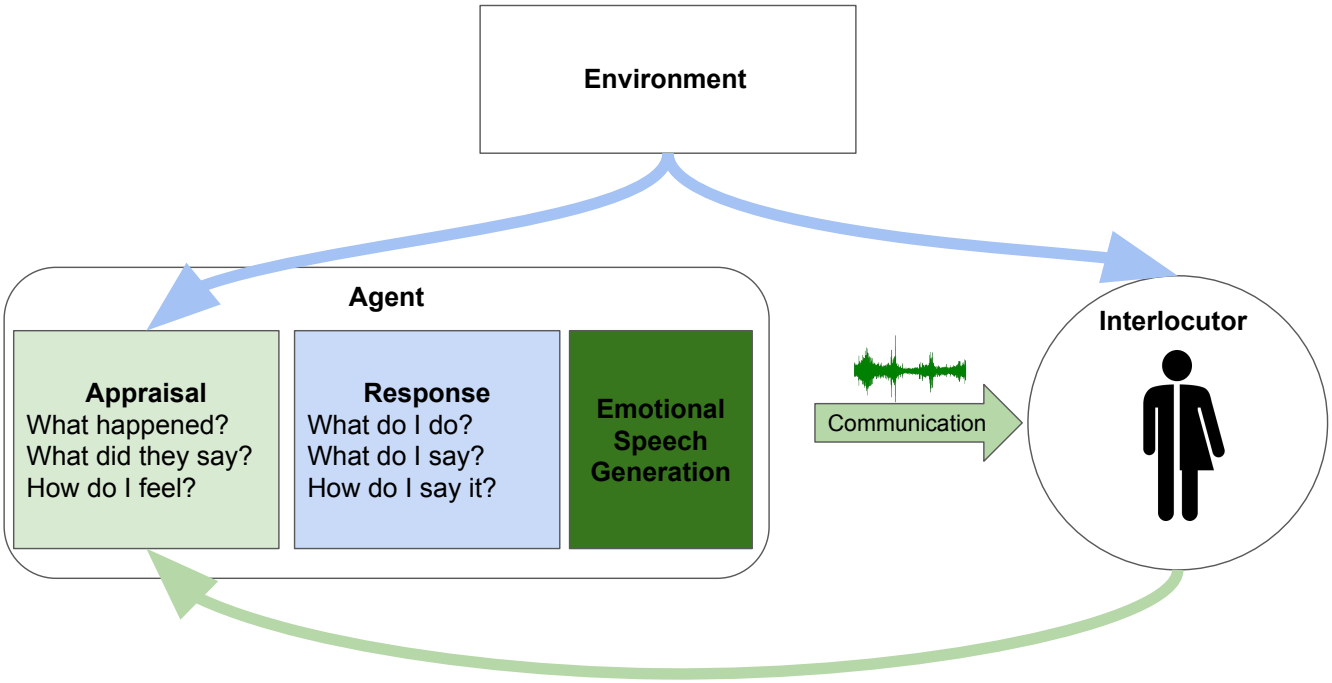


Fig. 1. Overview of our emotional speech generation model. We assume the presence of an artificial agent who can receive inputs from the environment (including responses from their interlocutor) and proceed to appraise the situation (an appraisal can be either hardcoded or learnt) and generate an appropriate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

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) [20]. 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 [26, 27], inputs from the environment are evaluated with respect to the agent's goals and concerns along several dimensions. For example, Ellsworth and Scherer [27] 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 it captures a core component of an affective agent: namely, that the appropriate text and emotion to be synthesised have to be somehow defined.

Of course, in most contemporary affective computing applications, the appropriate response is dictated by the creator of the application. Most of them contain hardcoded behaviours that the agents must follow (e. g., be constantly 'happy' or 'pleasing'). However, some recent works are already experimenting with learnt behaviours – for example, using reinforcement learning to train a dialogue agent to incorporate emotional responses [28], as these have been shown to increase subjective scores of dialogue richness [29, 30, 31]. As the field progresses, we expect more research towards less hardcoded and more learnt (or emerging) behaviours.

C. Emotional speech generation

For the present, we focus on the last step of an affective speech generation agent, which is the synthesis of the speech utterance itself, after a suitable text and emotion response have been determined by other processes [31]. An overview of this process is shown in Figure 2, 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 linguistic 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 began to erode with the advent of DL, with a single architecture often subsuming several (or even all) steps. We review both traditional, multi-step synthesis and DL-based synthesis in Section III.

Incorporating emotion to this pipeline is primarily done in two ways (green boxes and lines): Either an emotional voice conversion (EVC) module is tasked with adapting the emotion of the synthesised 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

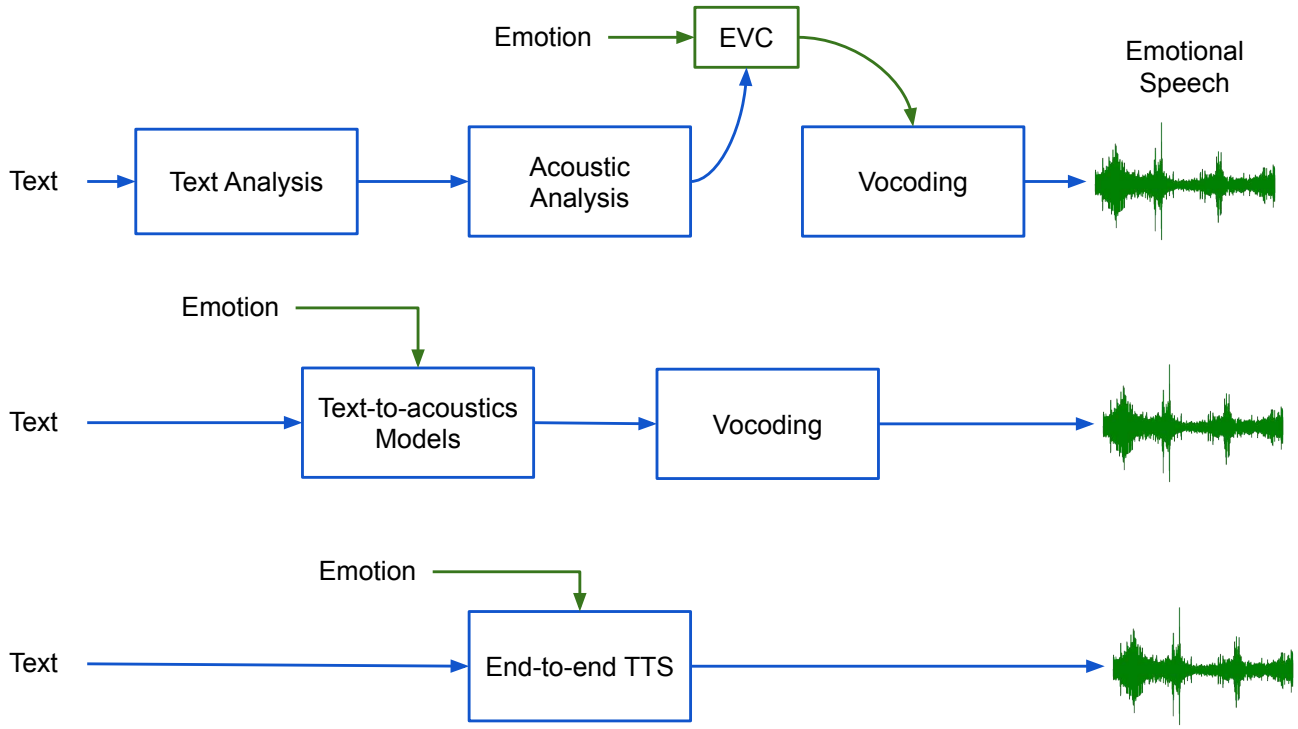


Fig. 2. Overview of an emotional speech synthesis module. Emotional synthesis (green) is superimposed on TTS pipelines (blue), which traditionally consist of 3 steps (top): text analysis, acoustic analysis, vocoding (synthesis). In the standard setup, emotion is used to modulate the acoustic features 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.

EVC; however, there are several works which 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 emotion may influence the generation of the text response itself, but as previously mentioned, we ignore this step for our purposes.

D. Computational models of affect

As Fehr and Russell [32] 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 [33]. Two of those have dominated the computational modelling of emotion [13]: discrete emotion theories, where emotions are considered to fall under discrete categories like Ekman's big six [34], and dimensional ones, like Russel's arousal, valence, and dominance [35]. 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 [36, 37]. Our review will accordingly focus on categorical ESS methods, as these have dominated the ongoing DL era.

E. Acoustic correlates of affect

Affective speech synthesis is concerned with adapting those constituents of a speech signal that convey 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, 38, 39]. 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 speech to suprasegmental features like prosody, voice quality (e. g., jitter and shimmer), spectral and energy features, and temporal patterns such as tempo and pausing [38, 39]. 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 [38]. 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 learnt representations, in the hope that those are better able to capture emotional information [40, 41]. 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 [42, 43, 44, 45, 46]. These approaches therefore constitute the main focus of our review.

There has also been some interest in other aspects of

vocalisations impacted by affect. For example, Tahon *et al.* [47] investigated the potential of generating emotional pronunciations to improve expressivity. More recently, Baird *et al.* [48] launched the Expressive Vocalisations Workshop and Competition (ExVo) to foster more research in the generation of realistic emotional "vocal bursts" [49]. 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 utterance 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, that, when properly combined, allowed visitors in his lab to recognise certain words [50]. 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 the next section. This section is concluded with an overview of (deep) voice conversion, an application field of voice transformation which has a lot in common with ESS [5].

A. 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* [51]. This paradigm was met with severe challenges in the modelling of articulatory behaviour, and was abandoned for a simpler, source-filter model that lends itself better to parameter control: *formant synthesis* [52, 53]. 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 favour 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 pre-recorded building blocks: words, syllables, half-syllables, phonemes, diphones, or triphones [54]. These pre-recorded units are concatenated to produce the utterance of interest. However, concatenative synthesis suffers from discontinuity effects (as the prosody of each recording can differ), and a dramatic increase in the data needed to cover all combinations of units.

All these downsides led to the adoption of a learning paradigm in the name of *statistical parametric speech synthesis (SPSS)* [55, 56]. SPSS adopts the three-stage model presented in Figure 2, 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 pre-processing steps (text normalisation, grapheme-to-phoneme conversion, etc.) followed by the extraction of all relevant features, like 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 excelled at this type of modelling [55]. Finally, the acoustic features are propagated to a suitable vocoder for the synthesis step. Some notable vocoders are WORLD [57] and STRAIGHT [58]. It is important to emphasise 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 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 centred around substituting HMMs with sequential models (RNNs [59] or long short-term memory networks (LSTMs) [60]) for acoustic modelling, such as the early DeepVoice systems [61, 62]. 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, like Tacotron [63, 64]. Nevertheless, several DL-based methods are still using the traditional 3-step pipeline, but substitute intermediate steps with their DL counterparts. The defining characteristics of deep speech synthesis are thus threefold: a) methods follow the SPSS formulation, b) all methods utilise deep neural networks in some, or all, steps of their pipeline, and c) some methods subsume some or all of the intermediate step in a single model – these are the so-called end-to-end approaches.

DL-based methods can be taxonomised along several categories:

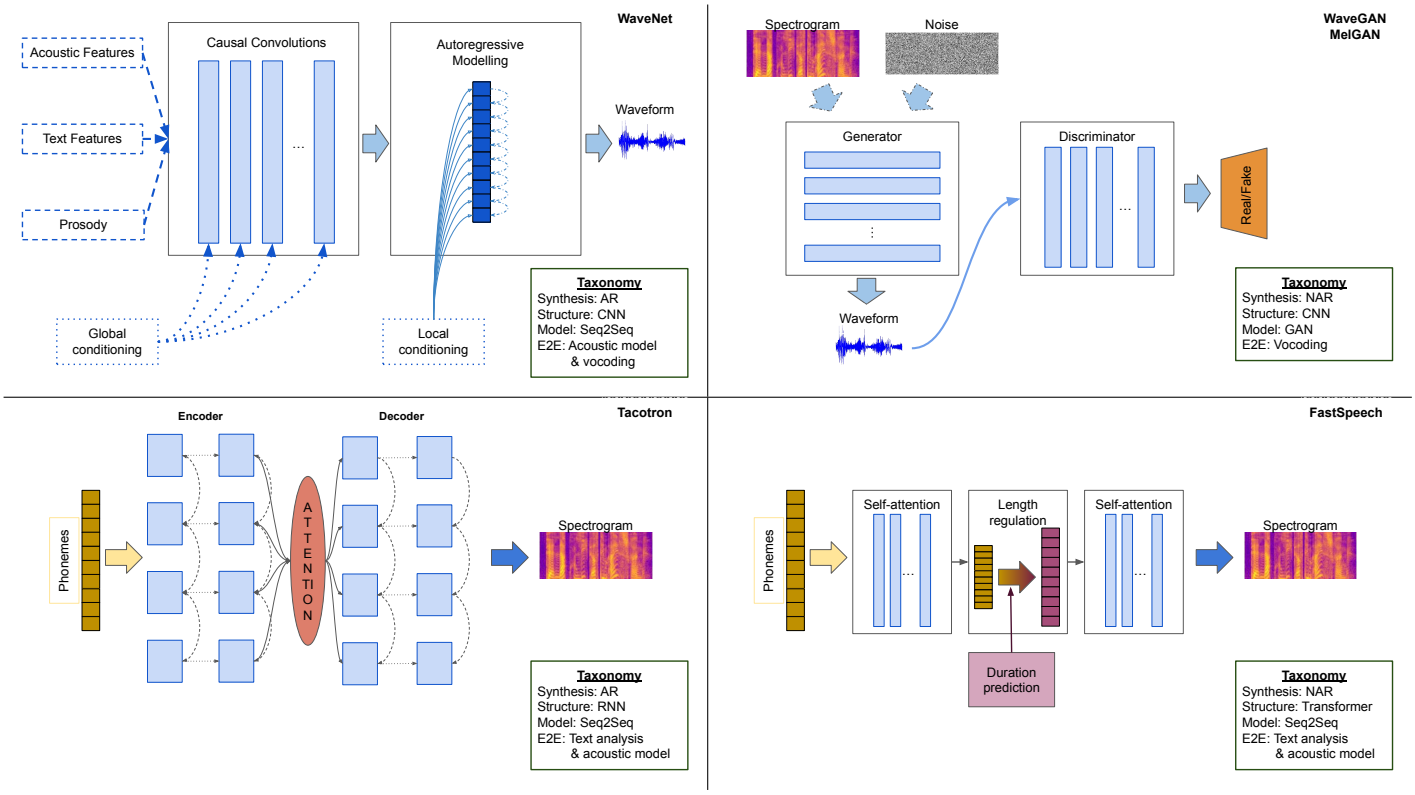


Fig. 3. Overview of main deep TTS paradigms. WaveNet was first introduced as a text-to-waveform model (thus combining an acoustic model with vocoding), which could be locally and globally conditioned on additional information; it was later extended to synthesise waveforms from input spectrograms, thus relegated to the role of a traditional vocoder. 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 as well as the mechanism to decide what text to output. Tacotron utilises 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 RNNs with Transformers.

- **Autoregressive** (AR) [14] vs **non-autoregressive** (NAR) structures [65].
- Type of **network structure**, where we primarily differentiated between convolutional neural networks (CNNs) [61, 62], sequential models (RNNs, gated recurrent unit networks (GRUs), LSTMs) [63, 64, 66] which may or may not include attention, and self-attention models (i. e., Transformers) [67, 68, 69].
- Type of **generative model** (e. g., variational autoencoder (VAE) [70], GAN [71]).
- **Degree of end-to-end** behaviour, which is characterised by the steps of the traditional SPSS pipeline that one or more (jointly trained) models subsume.

While such a categorisation 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 autoregressive 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 non-autoregressive synthesis [72] and changed to produce speech conditioned on (Mel-)spectrograms rather than linguistic features [66]. These rapid changes are expected as researchers continually optimise their pipelines in their quest for end-to-end synthesis. Nevertheless, as our focus is on presenting the core ideas that have revolutionised the TTS field in the last decade, we will primarily categorise approaches based on their earliest iterations.

An overview of recent, key TTS contributions from the deep learning era is shown in Figure 3. As previously mentioned, WaveNet [14] was the first neural model to be proposed for speech synthesis. In its first introduction, it was conceptualised 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 modelling: a) The use of dilated convolutions, which allowed it to increase its receptive field and model long-range interactions. b) 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 [66], thus effectively transforming it into a more traditional vocoder.

Tacotron [63] approached neural TTS by combining the two frontends of the SPSS pipeline, text analysis and acoustic modelling, 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 [73] whereas Tacotron2 used WaveNet [66]. Due to the sequential

nature of the encoder and 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 modelling such dependencies [74], and their ability to generate their output in non-autoregressive fashion by processing their inputs in parallel, Transformers were introduced as an alternative to RNNs in the FastSpeech series [68, 69]. 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 learnt 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 learnt 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 Goodfellow *et al.* [71], 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 [75], Parallel WaveGAN [76], HiFi-GAN [77] and others.

Sticking closer to the original formulation by Goodfellow *et al.* [71], the second category includes models like WaveGAN [78], 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 of models that can decide for themselves what they want to say.

We end this section 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-based [79] and diffusion-based models [80]. 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

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 voice conversion (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 sub-field 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 Mohammadi and Kain [81] and Sisman *et al.* [82].

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 [82, 83]; 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: Supra-segmental features like prosody and segmental ones like spectrum and formants. Short-time spectral features are correlates of *timbre*, which captures the 'tone' of an utterance and is related to the physiological characteristics of the speaker [83]. Prosody captures both physiological characteristics and speaking style [83]. 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 [83].

VC research has a history of more than 30 years. Early approaches utilised articulatory synthesis, but synthesised speech using the parameters of the target speaker [84]. More recent attempts used Gaussian mixture models (GMMs) [85], exemplar-based frameworks based on non-negative matrix factorisation (NMF) [86], and HMMs [87]. However, despite several attempts, no notable progress was made until recent years, which saw advent of deep neural networks (DNNs). Recently proposed methods exploit the representation power of DNNs by means of VAEs [88, 89], GANs [90, 91], and seq2seq models [92, 93].

A key distinction of VC approaches is between those who use *parallel* and *non-parallel* 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 means 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 non-parallel 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 the previous section, WaveNet supports both global and local conditioning – these conditioning interfaces can be co-opted by voice conversion algorithms to change supra-segmental 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 [89], d-vectors [88], or bottleneck features as speaker representations from a DNN [94].

As with TTS, GANs also play a prominant 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 non-parallel data, an additional form of regularisation is proposed by the introduction of a cycle-consistency loss [95], resulting in CycleGAN [91, 96]. 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 utilised 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 [90, 97], which extends the consistently principle to multiple source-target domains. Finally, seq2seq models, which use an encoder-decoder architecture, have also been extensively studied in the VC field [92, 93]. 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

Emotional speech synthesis (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 sub-component, as generating emotional speech does in fact entail the generation of non-emotional speech, meaning utterances that are comprehensible for their intended linguistic and emotional meaning. However, most contemporary works envision ESS (or, 'expressive' TTS as they often name it) 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. As for TTS, we start with a brief history of earlier works.

A. A 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 synthesisers that took care of the synthesis. The values of these parameters were usually chosen by experts, and their suitability verified by follow-up recognition studies. A more data-driven study by Burkhardt and Sendlmeier [98] investigated instead different parameter ranges and identified those that lead to better recognition rates, rather than setting them a priori, but nevertheless relied on customly manipulating these parameters for synthesis.

While rule-based ESS brought some initial excitement to the field, it was later abandoned in favour of concatenative synthesis [16]. Like its TTS counterpart, concatenative ESS relied on selecting speech units uttered with the appropriate emotions from an existing dataset. As a result, it too suffered from the same shortcomings that plagued standard speech synthesis, namely the lack of available data and discontinuities.

Finally, following a similar trend as TTS, ESS transitioned to a data-driven paradigm with the advent of SPSS [99, 100], 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 synthesise speech. In particular, mappings between both prosodic and spectral features were learnt using data [99, 100]. As is the case for voice conversion, this entailed the presence of *parallel data* from whence the mappings can be learnt.

B. Taxonomy of deep ESS approaches

There are various ways in which to taxonomise deep ESS approaches, as shown in Figure 4. The first one is based on whether they perform **text-to-emotional-features synthesis (TTEF)**, where they go directly from text to emotionally-laden acoustic features, or **emotional voice conversion (EVC)**, where they rely on acoustic features which are already generated by a standard TTS system. Due to the widespread success of TTS, most ESS systems are essentially performing EVC, as they utilise existing components which 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 synthesising one from the ground up. This allows EVC to be added as an extra step in a traditional TTS pipeline – first synthesise, and then convert to the target emotion. This decomposition into two constituents can reduce computational complexity and dependence on data.

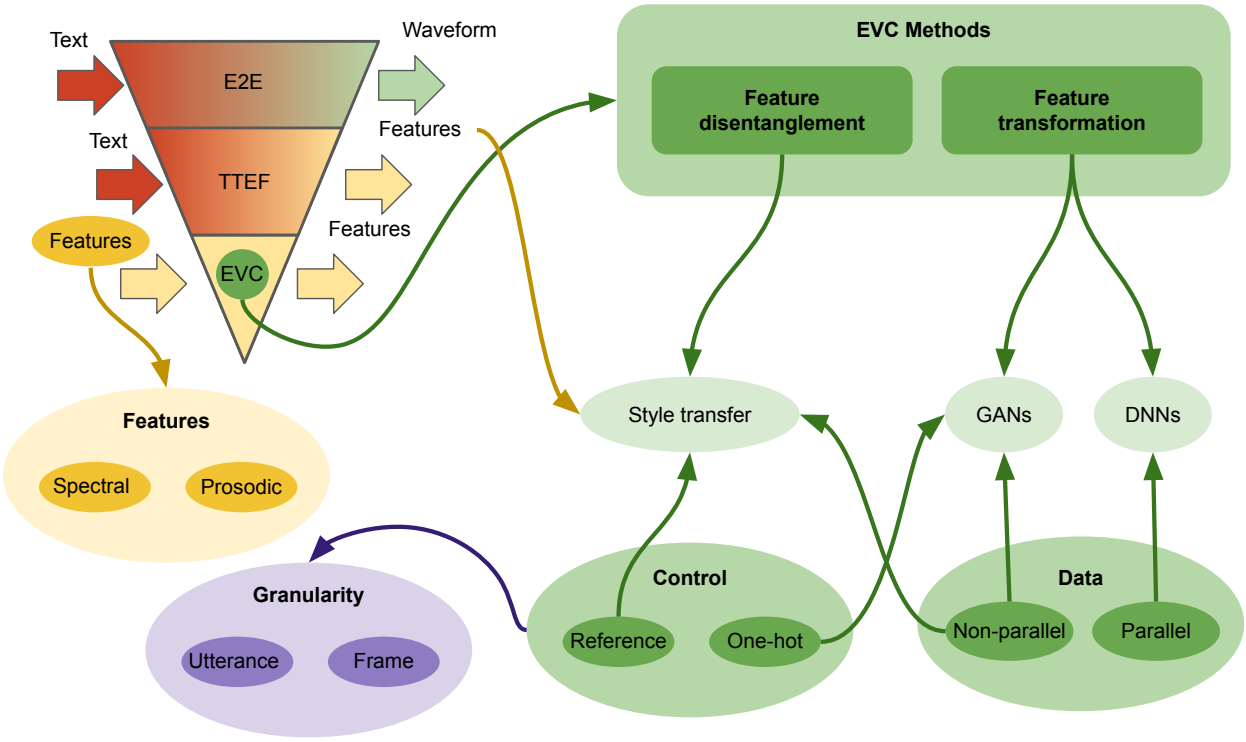


Fig. 4. Taxonomy of deep emotional speech synthesis approaches. Approaches can be primarily differentiated according to the following ways: (a) How many steps of the synthesis they incorporate, which is in term determined by their input and output, accordingly categorised as end-to-end (E2E), text-to-emotional-features (TTEF), or emotional voice conversion (EVC) methods. (b) How control is achieved, as well as the level of granularity that this control can achieve. (c) For EVC methods, on whether they use parallel or non-parallel data. (d) For non-parallel 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.

Furthermore, vocoding is to 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 they require for training. Approaches relying on **parallel data** learn a mapping 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 which can work on **non-parallel 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 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 it needs to produce. The latter provides instead a non-auditory representation of emotions, with the choice of representation being a sub-category 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 learnt by DNNs) that can be used to increase the span of covered emotions. This is also related to the type of **features** which 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

[4]One exception is Kreuk *et al.* [115], which use the emotion label to condition a HiFi-GAN vocoder.

TABLE I
MAJOR DL-BASED EMOTIONAL SPEECH SYNTHESIS WORKS CATEGORISED ACCORDING TO THE TAXONOMY PRESENTED IN FIGURE 4.

Approach	Control	Intensity	Granularity	Non-parallel data	Conversion	Model	Features	End-to-end
Ming *et al.* [101]	Fixed	N/A	Utterance	✗	Transformation	bLSTM	STRAIGHT	EVC
Lee *et al.* [102]	One-hot	N/A	Utterance	✗	N/A	Seq2Seq	Spectra	TTEF
Lorenzo-Trueba *et al.* [42]	Annotator agreement	Annotator agreement	Utterance	✗	N/A	RNN	WORLD	TTEF
Choi *et al.* [103]	Reference	N/A	Utterance	✗	Disentanglement	CNN	Spectra	TTEF
Kwon *et al.* [104]	Reference	N/A	Utterance	✗	Disentanglement	Seq2Seq	Spectra	TTEF
Shankar *et al.* [43]	Fixed	N/A	Utterance	✗	Transformation	Highway	F0/intensity	EVC
Bao *et al.* [105]	Fixed	N/A	Utterance	✓	Transformation	CycleGAN	openSMILE	EVC
Luo *et al.* [44]	Fixed	N/A	Utterance	✗	Transformation	GAN	F0	EVC
Robinson *et al.* [106]	Fixed	N/A	Frame	✗	Transformation	Seq2Seq	F0	EVC
Gao *et al.* [107]	Reference	N/A	Utterance	✓	Disentanglement	GAN	F0/Spectra	EVC
Kim *et al.* [108]	Reference	N/A	Utterance	✗	Disentanglement	Seq2Seq	Spectra	EVC
Rizos *et al.* [109]	One-hot	N/A	Utterance	✓	Transformation	StarGAN	Cepstra	EVC
Cao *et al.* [110]	Fixed	N/A	Utterance	✓	Transformation	VAE-GAN	Cepstra	EVC
Schnell and Garner [45]	Reference	Saliency maps	Frame	✗	Transformation	RNN	WORLD	EVC
Liu *et al.* [111]	Reference	N/A	Utterance	✓	Transformation	Seq2Seq	Spectra	TTEF
Du *et al.* [112]	Reference	N/A	Utterance	✓	Transformation	StarGAN	Cepstra	EVC
Choi and Hahn [40]	Reference	Manual	Utterance	✗	Disentanglement	Seq2Seq	Spectra	EVC
Cai *et al.* [113]	Reference	N/A	Utterance	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Wu *et al.* [114]	Reference	N/A	Frame	✗	Disentanglement	Seq2Seq	Spectra	TTEF
Kreuk *et al.* [115]	Fixed	N/A	Frame	✗	Transformation	Seq2Seq	Spectra/F0/T	EVC
Zhou *et al.* [41]	Reference	Ranking	Utterance	✓	Disentanglement	Seq2Seq	Spectra	EVC
Zhang *et al.* [46]	Reference	Posterior	Utterance	✓	Disentanglement	Seq2Seq	Spectra/F0	EVC
Li *et al.* [116]	Reference	Manual	Utterance	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Liu *et al.* [25]	Reference	N/A	Frame	✓	Disentanglement	Seq2Seq	Spectra	TTEF
Lei *et al.* [117]	Reference	Ranking	Frame	✓	Disentanglement	Seq2Seq	Spectra	TTEF

that are modified. The type 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 control requires more fine-grained approaches.

Finally, as for TTS systems, there are different degrees in which approaches transition from a multi-step (or 'cascade') SPSS paradigm, where any one step can be implemented via a DNN, towards an end-to-end 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 inherit all its properties, like the type of architecture or underlying model, as discussed in Section III.

We have categorised major DL-based ESS works according to our taxonomy in Table I. Whenever researchers experimented with more than one disjunct categories in their work, we chose to assign them according to where their major focus lay. In the following subsections, we proceed to analyse the discrete categories and refer to these works in detail.

C. 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 [42, 104, 108] – 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 utilised to learn the mapping between acoustic features, or even directly from 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 generalise 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.

D. Non-parallel data methods

Transitioning to methods capable of handling non-parallel data was a necessary prerequisite for the development of more naturalistic and generalisable 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 be processed. In particular, there was now no matching sentence of the target emotion for the text that needed to be synthesised (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 the next two subsections, this decomposition could be implicitly enforced to the model via manipulation of its training strategy, leading to a set of methods we name *direct feature transformation methods*, or explicitly designed into it, leading to *disentanglement methods*. In either case, the increased complexity of handling non-parallel 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 I.

E. 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:

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

where x_s and x_t are samples sourced from X_S and X_T, respectively. As we saw in Section IV-C, 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, non-parallel 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 are generative adversarial networks (GANs). GANs were first introduced by Goodfellow *et al.* [71], 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 follows 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 formalised as follows:

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{x\sim p(x)}[logD(x)] \\ &+ \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))],\end{aligned} \tag{2}$$

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 follows 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 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 speaker or linguistic content. As a consequence, the standard GAN paradigm needed to be modified for non-parallel training data.

One of the most prominent approaches that deal with the aforementioned problem is the cycle-consistent generative adversarial network (CycleGAN) [95]. 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:

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

where $G_{S\rightarrow T}$ and $G_{T\rightarrow S}$ are the generators of the two GANs, and x_s and x_t are data from the domains X_S and X_T,

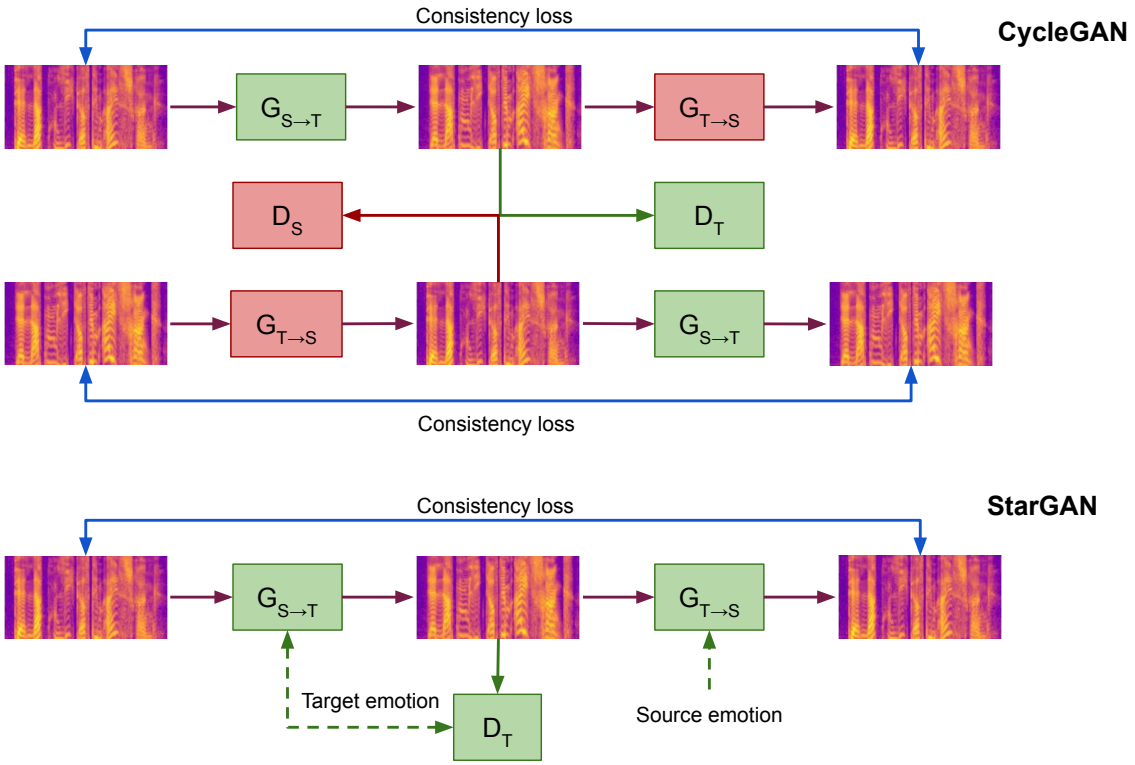


Fig. 5. Overview of the two main GAN paradigms for emotional voice conversion. CycleGAN (top) first converts the source utterance to the target domain, which is then evaluated by the discriminator, while another generator maps it back to its source domain to compute the consistency loss. The inverse process is followed to map from the target domain to the source domain which helps to regularise 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.

respectively. The full loss function of a CycleGAN is given as:

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

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

CycleGAN was first transferred to the speech domain by Kaneko and Kameoka [91], who used it to perform voice conversion. In order to do so, they enhanced the generator networks with *gated CNNs* [118]. Additionally, they implemented an *identity loss* [119], which ensures that samples already belonging to the target domain are not altered. They formulate the identity loss as:

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

A more sophisticated version of their work that introduces additional discriminator networks that are applied on the circularly converted voice was presented in Kaneko *et al.* [96]. Later, EVC approaches then adopted this formulation for their purposes [105, 120, 121, 122].

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 [97, 109, 123]. The training concepts of both models are illustrated in Figure 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 as follows:

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

Additionally, 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 with the discriminator and generator to distinguish between the different domains. This classifier is used to build the StarGAN's classification loss component:

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

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

F. 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 form 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 one related to emotion.

The basic schema, as followed by recent works [41, 46, 113], is shown in Figure 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; 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 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 [25, 41, 116]. This way, they are guided to learn emotional information.

Inspired by advances in TTS, with models like Tacotron achieving impressive synthesis results, recent works have transitioned to conditioning such models on emotional information [25, 113, 114, 116, 117], thus moving a step up from EVC methods in the end-to-end hierarchy and going directly from 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 for the content part.

G. Controlling the synthesised 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 overall speaking 'style' (for example, read vs 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, inference was simply done by selecting the appropriate pair. This is essentially superseded by methods which use one-hot encodings of emotions, like StarGAN [109]; there each emotion is represented by a one-hot vector. 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 from recent advances in expressive TTS. The core proposition behind style transfer is to learn an encoding for those aspects of speech that correspond to emotion in a data-driven way. This idea was first introduced in Skerry-Ryan *et al.* [124]. The authors utilised 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 learnt; 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 learnt 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 [124].

To further promote disentanglement between different factors, Wang *et al.* [125] introduced global style tokens (GSTs). Utilising a similar setup as Skerry-Ryan *et al.* [124], 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 ran 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

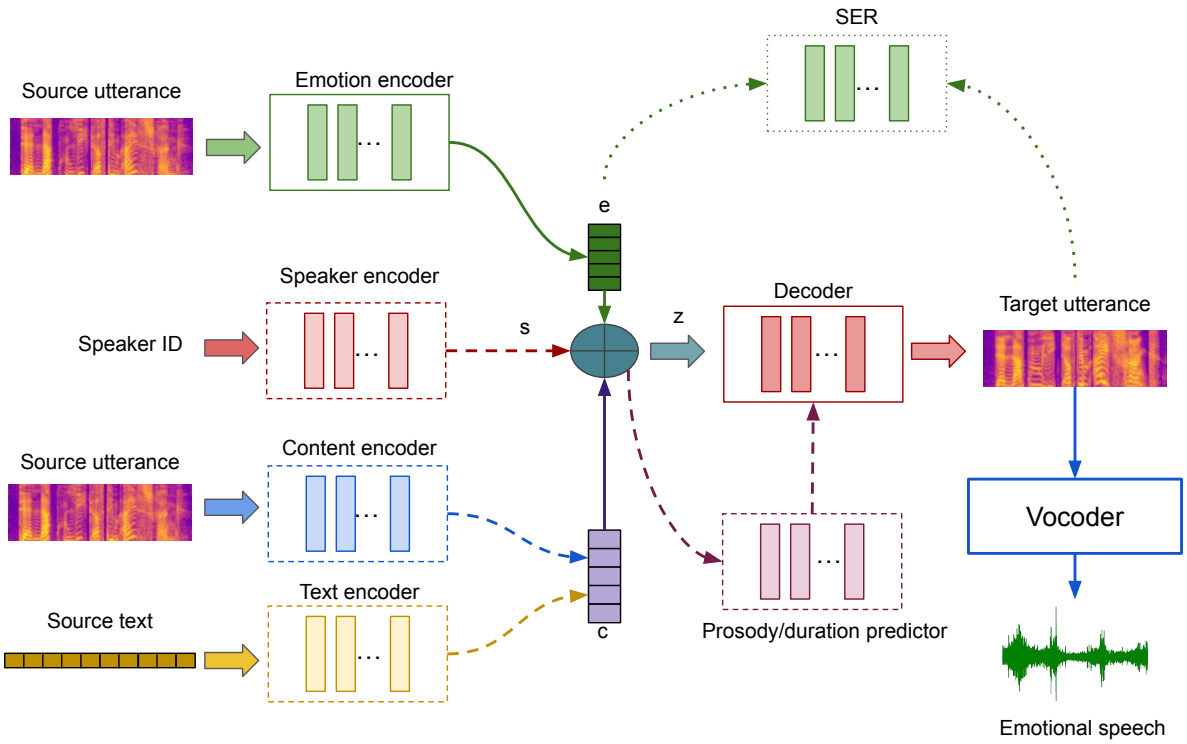


Fig. 6. Overview of disentanglement methods for emotional voice conversion. The source utterance is passed through an emotion encoder to generate an emotion embedding. Content information is additionally provided from encoders operating on the acoustic or linguistic features – or both. Optionally, speaker embeddings are provided by a speaker encoder. All embeddings are fed to a decoder which generates the target utterance. This utterance, as well as the emotion embeddings, are often evaluated with respect to the emotion information they contain, usually by a pre-trained 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.

emotion during training [25, 113, 116, 126]. This already biases the reference encoder to capture the differences in emotion, and results in better controllability during inference as well. Moreover, it is often the case that an average style is computed on all training data and stored as reference for each category, therefore sparing the need to select an appropriate reference during inference.

H. 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 [49]. 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.* [42] 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 [40] and Li *et al.* [116].

While 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 in Schnell and Garner [45], who use the saliency maps of pretrained, attention-based SER systems as guidance for frame-level intensity control. Similarly, Zhang *et al.* [46] use the posterior probabilities of an SER for utterance-level emotional control. Both approaches follow a reference-based paradigm for inference-time control. Ultimately though, those approaches are limited by the effectiveness of the SER systems, which, although greatly improving in recent years [127], 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 [128]. Zhou *et al.* [41] and Lei *et al.* [117] 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 optimisation 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 synthesised emotion.

I. 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, as well as 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 [45] 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.* [114] achieve this with a capsule network [129] while Liu *et al.* [25], Kreuk *et al.* [115], and Lei *et al.* [117] achieve this via frame-level losses. As seen in Table I, this trend is picking up pace this last year with several very recent works pursuing higher degrees of granularity.

J. Features manipulated to achieve expressivity

As Table I shows, 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-E. Another factor which 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 [42, 45, 101, 106, 109], while works that only intended to use EVC as a means to improve SER performance [105] manipulated feature vectors used by the downstream models [130]. 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 [41, 102, 103, 104, 108, 121]. Manipulation of other features, such as F0, is still done but primarily using simple statistical techniques (e. g., by standardising the F0 curve with the statistics of the target emotions [109]) and rarely using deep learning methods [43, 115]. 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.

K. 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 neighbouring field allows the community to iterate quickly over problems that are specific to ESS rather than reinventing the wheel. As seen from Table I, the majority of EVC models are relying on seq2seq models [41, 45, 46, 102, 104, 106, 108, 111]. This is counter to earlier methods, which relied on highway networks or simpler sequential models [42, 43, 101, 103]. 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 length than those seen by the encoder [18]. A more thorough overview of seq2seq models for EVC can be found in Yang *et al.* [18]. This seq2seq trend is also followed by more recent TTEF methods, which rely on the Tacotron architecture [25, 113, 114, 116, 117] and merely condition it with emotional information. Methods using adversarial models like CycleGAN [105] and StarGAN [109] also stay close to their original versions, with minor adaptations to fit the EVC problem. Finally, some works attempt to leverage representations learnt by large, pre-trained models and thus rely on transformer-based architectures such as HuBERT [115].

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: (a) an encoder, which reduces the dimensions of the speech signal to a latent representation (or code), and (b) 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 realisation of AE that is sometimes used for EVC as well is the VAE [110]. 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 $\mathbf{z}{\sim}\mathcal{N}(\mathbf{z}; \mu, diag(\sigma^2))$. The encoder tries to estimate the mean and variance of the distribution, and with the use of the reparameterisation 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 maximising a variational lower bound of the log-likelihood:

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

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

L. 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". 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 judgement 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) [131]. 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 [132]. Some dimensions commonly used in recent works are naturalness [25, 40, 44, 107], speech quality [45, 112], emotional strength [42], and similarity with the target emotion [46, 103, 104, 114, 116]. ABX tests are also commonly used [44, 112], 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 [133]. 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 synthesised emotions [40, 45, 106, 110, 111, 113].

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 [132]. In the case of TTS, for example, BSD [134], PESQ [135], POLQA [136], or ITU-T Rec. P.563 [137] are often used to evaluate the quality of generated signals. No such standardised 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 pre-trained SER models to judge whether generated samples are correctly classified [138]. 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.

M. 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 Zhou *et al.* [17]. The authors mention five key desiderata for designing datasets that cover all conditions necessary for generalisation: 1) Increasing lexical variability, as emotional speech datasets are often recorded using a limited set of datasets. 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 generalise well to new speakers. 4) Controlling for confounders, such as different accents or demographics. 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 generalise 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 [42, 104, 108] or relied on the standard SER datasets, such as IEMOCAP [139], EMO-DB [140], etc.

N. 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 mean opinion score (MOS) scores of 4 for 6 [116, 117] or 7 [46] emotion categories certainly suggest that we are approaching a 'WaveNet moment' for ESS as well, as the revolution started by Oord *et al.* [14] began with such MOS scores for naturalness. Accordingly, some works are showing subjective emotion recognition accuracies reaching up to 80% [40, 111]. Other works, however, feature much lower scores, dropping down to almost 50% recognition accuracy for 4 emotions [107, 110]. 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: [116, 117] used read speech datasets recorded by single, female authors specifically constructed for ESS, while [107, 110] both used IEMOCAP [139] which includes improvised 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 they are expect to operate in. It depends on the number and kind of emotions the agent is expected to support, the languages and cultures it needs to cover, their malleability to user input, their robustness to different noise conditions, etc. As ESS make 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 all 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 the last subsection, we argued that while DL constitutes an immense leap forward compared to previous approaches, ESS remains far from solved. In this section, we highlight the main limitations, discuss whether the ESS problem should be solved at all given the ethical considerations it raises, and finally outline some promising areas of future research.

A. Main limitations

As seen in Section IV-L, ESS is still plagued by a lack of holistic, standardised 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 use similar datasets and report similar metrics.

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 [127], where models learn an entangled representation of linguistics and acoustics [141]. 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 automatic speech recognition (ASR), shows that algorithms will generalise 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 neighbouring 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 also 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) [142]. Therefore, it could be that the gap between humans and machines cannot be bridged until the latter also acquire 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 issue 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) [143]. 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 criticised in recent a UNESCO report [144], 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 behaviour 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 cognisant of the fact that they are interacting with an artificial entity under all circumstances [145], but the question remains if that is sufficient to mitigate the potential dangers that could arise from 'overhumanising' those entities.

The question of 'should' does, however, not cover the degree in 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 *generalisability*: Do we cover all different cultures? 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 synthesised samples, are typically annotated from 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 provides a solid background from which to start. Co-opting those approaches for the evaluation of synthesised emotional speech, and adopting best-practices from the sister-domain of TTS, seems 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 a transformative technology like ESS. However, specifically for the field of computational paralinguistics, the reader is referred to Batliner *et al.* [146].

C. Future perspectives

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

Synthesising 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 [49]. Synthesising 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 [147].

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 for several decades [148]. This topic has been recently revived in the context of big language models, which can be adapted to emulate a specific personality [149]. As personality has been also shown to manifest in speech signals [150], it is an evident next step to introduce it to conversational agents as well [151]. 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 modelled from speech: deception, sincerity, nativeness, cognitive load, likability, interest, and others are all variables which could be added to the capabilities of affective agents.

Personalisation is expected to be another major aspect of future ESS systems. Both the expression [152, 153, 154] and the perception [155] 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 [156]

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 grounds 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 utilised 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 [157]. An overture to this exciting domain can already be found in intelligent dialogue generation, where reinforcement learning is already being used to adjust the linguistic style of an agent [158] or to learn backchanneling responses [159, 160]. 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

[5]https://www.youtube.com/watch?v=yDI5oVn0RgM
[6]https://www.competitions.hume.ai/exvo2022
[7]www.compare.openaudio.eu

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 emotional 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 towards more 'end-to-end' emotional synthesis. Finally, following recent successes on conversion of one emotion category to another, albeit in the restricted domain of acted and read speech, several works are now focusing on the control of emotional 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, the overrepresentation of a few cultures and languages in 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 its 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 realise the full potential of affective agents.

ACKNOWLEDGMENT

This work has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

VII. REFERENCES

[1] A. J. Fiannaca, A. Paradiso, J. Campbell, and M. R. Morris, "Voicesetting: Voice authoring UIs for improved expressivity in augmentative communication," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Montreal, Canada: ACM, 2018, pp. 1–12.

[2] M. Minsky, *Society of mind*. Simon and Schuster, 1988.

[3] R. W. Picard, *Affective computing*. 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Taipei, Taiwan, 2009, pp. 3585–3588.

[6] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[7] I. R. Murray, "Simulating emotion in synthetic speech.," Ph.D. dissertation, University of Dundee, 1989.

[8] J. E. Cahn, "Generating expression in synthesized speech," Ph.D. dissertation, Massachusetts Institute of Technology, 1989.

[9] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American voice I/O society*, vol. 8, no. 1, pp. 1–1, 1990.

[10] Y. Kitahara and Y. Tohkura, "Prosodic control to express emotions for man-machine speech interaction," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 75, no. 2, pp. 155–163, 1992.

[11] B. Granström, "The use of speech synthesis in exploring different speaking styles," *Speech Communication*, vol. 11, no. 4-5, pp. 347–355, 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," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

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

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 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," *arXiv preprint arXiv:2106.15561*, 2021.

[16] M. Schröder, "Emotional speech synthesis: A review," *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 561–564, 2001.

[17] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 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 *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Incheon, South Korea: ISCA, 2022.

[19] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014.

[20] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2011.

[21] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, ONT, Canada, 2021, pp. 6593–6597.

[22] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[23] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brno, Czech Republic: ISCA, 2021, pp. 3616–3619.

[24] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. A. Nguyen, M. Riviere, A. Mohamed, E. Dupoux, *et al.*, "Text-free prosody-aware generative spoken language modeling," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: ACL, 2022, pp. 8666–8681.

[25] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[26] K. R. Scherer, "Appraisal theory," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power, Eds., John Wiley & Sons Ltd, 1999, pp. 637–663.

[27] P. C. Ellsworth and K. R. Scherer, *Appraisal processes in emotion.* Oxford University Press, 2003.

[28] R. Lan, J. Wang, W. Huang, Z. Deng, X. Sun, Z. Chen, and X. Luo, "Chinese emotional dialogue response generation via reinforcement learning," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 4, pp. 1–17, 2021.

[29] 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 *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, Melbourn, Australia: ACL, 2018, pp. 371–375.

[30] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, New Orleans, LA, USA: AAAI, 2018.

[31] Z. Bucinca, Y. Yemez, E. Erzin, and M. Sezgin, "AffectON: Incorporating affect into dialog generation," *IEEE Transactions on Affective Computing*, 2020. DOI: 10.1109/TAFFC.2020.3043067.

[32] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective," *Journal of Experimental Psychology: General*, vol. 113, no. 3, p. 464, 1984.

[33] K. R. Scherer, "Towards a prediction and data driven computational process model of emotion," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 279–292, 2019.

[34] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[35] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[36] M. Schroder, "Expressing degree of activation in synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1128–1136, 2006.

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

[38] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, p. 614, 1996.

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

[40] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.

[41] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, pp. 1–18, 2022. DOI: 10.1109/TAFFC.2022.3175578.

[42] 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 Communications*, vol. 99, pp. 135–143, 2018.

[43] R. Shankar, J. Sager, and A. Venkataraman, "A multi-speaker emotion morphing model using highway networks and maximum likelihood objective," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria: ISCA, 2019, pp. 2848–2852.

[44] 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 Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[45] B. Schnell and P. N. Garner, "Improving emotional TTS with an emotion intensity input from unsupervised extraction," in *Proceedings 11th ISCA Speech Synthesis Workshop*, Brno, Czech Republic: ISCA, 2021, pp. 60–65.

[46] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre," *arXiv preprint arXiv:2206.14866*, 2022.

[47] M. Tahon, G. Lecorvé, and D. Lolive, "Can we generate emotional pronunciations for expressive speech synthesis?" *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 684–695, 2018.

[48] A. Baird, P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen, "The ICML 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts," *arXiv preprint arXiv:2205.01780*, 2022.

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

[50] H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgang von Kempelen," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 151–166, 1950.

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

[52] J. Allen, S. Hunnicutt, R. Carlson, and B. Granstrom, "MITalk-79: The 1979 MIT text-to-speech system," *The Journal of the Acoustical Society of America*, vol. 65, no. S1, S130–S130, 1979.

[53] D. H. Klatt, "Review of text-to-speech conversion for english," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[54] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[55] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, vol. 3, Istanbul, Turkey, 2000, pp. 1315–1318.

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

[57] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[58] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

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

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

[61] S. Ö. Arik, M. Chrzanowski, A. Coates, G. F. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Y. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, D. Precup and Y. W. Teh, Eds., vol. 70, Sydney, NSW, Australia: PMLR, 2017, pp. 195–204.

[62] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[63] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden: ISCA, 2017, pp. 4006–4010.

[64] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R.-S. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada: IEEE, 2018, pp. 4779–4783.

[65] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, Vienna, Austria, 2020, pp. 7586–7598.

[66] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Calgary, AB, Canada, 2018, pp. 4779–4783.

[67] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA: AAAI, 2019, pp. 6706–6713.

[68] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[69] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia: PMLR, 2020.

[70] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA: PMLR, 2013.

[71] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[72] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, Stockholm, Sweden, 2018, pp. 3918–3926.

[73] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[75] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2019, pp. 14 881–14 892.

[76] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, 2020, pp. 6199–6203.

[77] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[78] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada: PMLR, 2018.

[79] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, United Kingdom, 2019, pp. 3617–3621.

[80] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, Baltimore, MD, USA, 2021, pp. 8599–8608.

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

[82] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

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

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

[85] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[86] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, B. Yegnanarayana, Ed., Hyberabad, India: ISCA, 2018, pp. 1978–1982.

[87] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France: IEEE, 2006, pp. 77–80.

[88] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada: IEEE, 2018, pp. 5274–5278.

[89] 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 *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, South Korea: IEEE, 2016, pp. 1–6.

[90] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, Stuttgart, Germany: IEEE, 2018, pp. 266–273.

[91] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proceedings of the IEEE European Signal Processing Conference (EUSIPCO)*, Rome, Italy: IEEE, 2018, pp. 2100–2104.

[92] S. Liu, Y. Cao, S. Kang, N. Hu, X. Liu, D. Su, D. Yu, and H. Meng, "Transferring source style in non-parallel voice conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, H. Meng, B. Xu, and T. F. Zheng, Eds., Shanghai, China: ISCA, 2020, pp. 4721–4725.

[93] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, 2019, pp. 6805–6809.

[94] 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 *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Honolulu, HI, USA: IEEE, 2018, pp. 829–833.

[95] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy: IEEE, 2017, pp. 2242–2251.

[96] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, 2019, pp. 6820–6824.

[97] Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," Salt Lake City, UT, USA: IEEE, 2018, pp. 8789–8797.

[98] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant synthesis," in *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*, Beijing, China: ISCA, 2000.

[99] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *Proceedings of the International Conference on Speech Prosody*, Nara, Japan: ISCA, 2004, pp. 413–416.

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

[101] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, N. Morgan, Ed., San Francisco, CA, USA: ISCA, 2016, pp. 2453–2457.

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

[103] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, 2019, pp. 6950–6954.

[104] O. Kwon, I. Jang, C. Ahn, and H. Kang, "Emotional speech synthesis based on style embedded Tacotron2 framework," in *Proceedings of the International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, Jeju, South Korea, 2019, pp. 1–4.

[105] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, G. Kubin and Z. Kacic, Eds., Graz, Austria: ISCA, 2019, pp. 2828–2832.

[106] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of F0 for speech emotion conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, United Kingdom: IEEE, 2019, pp. 6830–6834.

[107] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, G. Kubin and Z. Kacic, Eds., Graz, Austria: ISCA, 2019, pp. 2858–2862.

[108] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Barcelona, Spain, 2020, pp. 7774–7778.

[109] G. Rizos, A. Baird, M. Elliott, and B. W. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proceedings of the IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, 2020, pp. 3502–3506.

[110] Y. Cao, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "Nonparallel emotional speech conversion using VAE-GAN," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Shanghai, China: ISCA, 2020, pp. 3406–3410.

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

[112] Z. Du, B. Sisman, K. Zhou, and H. Li, "Expressive voice conversion: A joint framework for speaker identity and emotional style transfer," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, Cartagena, Colombia, 2021, pp. 594–601.

[113] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, ONT, Canada, 2021, pp. 5734–5738.

[114] X. Wu, Y. Cao, H. Lu, S. Liu, S. Kang, Z. Wu, X. Liu, and H. Meng, "Exemplar-based emotive speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 874–886, 2021.

[115] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T.-A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using decomposed and discrete representations," *arXiv preprint arXiv:2111.07402*, 2021.

[116] 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 Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448–1460, 2022.

[117] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[118] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, PMLR, Sydney, NSW, Australia, 2017, pp. 933–941.

[119] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France: PMLR, 2017.

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

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

[122] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "An improved cycleganbased emotional voice conversion model by augmenting temporal dependency with a transformer," *Speech Communication*, 2022.

[123] X. He, J. Chen, G. Rizos, and B. Schuller, "An improved StarGAN for emotional voice conversion: Enhancing voice quality and data augmentation," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brno, Czech Republic: ISCA, 2021, pp. 821–825.

[124] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden: PMLR, 2018, pp. 4693–4702.

[125] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden: PMLR, 2018, pp. 5180–5189.

[126] P.-F. Wu, Z.-H. Ling, L.-J. Liu, Y. Jiang, H.-C. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Lanzhou, China: IEEE, 2019, pp. 623–627.

[127] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[128] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, San Antonio, TX, USA, 2017, pp. 248–255.

[129] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[130] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, Florence, Italy: ACM, 2010, pp. 1459–1462.

[131] S. Robert C., W. Stefan, and H. David S., "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, pp. 213–227, 2016.

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

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

[134] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[135] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

[136] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[137] L. Malfait, J. Berger, and M. Kastner, "P.563 - the ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[138] A. Baird, S. Mertes, M. Milling, L. Stappen, T. Wiest, E. André, and B. W. Schuller, "A prototypical network approach for evaluating generated emotional speech," Brno, Czech Republic: ISCA, 2021, pp. 3161–3165.

[139] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[140] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of German emotional speech.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. 5, Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[141] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

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

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

[144] M. West, R. Kraut, and H. Ei Chew, "I'd blush if i could: Closing gender divides in digital skills through education," 2019.

[145] E. Commission, "White paper on artificial intelligence: A european approach to excellence and trust," *Com (2020) 65 Final*, 2020.

[146] A. Batliner, S. Hantke, and B. Schuller, "Ethics and Good Practice in Computational Paralinguistics," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1236–1253, 2022.

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

[148] B. L. Brown, W. J. Strong, and A. C. Rencher, "Perceptions of personality from speech: Effects of manipulations of acoustical parameters," *The Journal of the Acoustical Society of America*, vol. 54, no. 1, pp. 29–35, 1973.

[149] S. S. Keh, I. Cheng, *et al.*, "Myers-briggs personality classification and personality-specific language generation using pre-trained language models," *arXiv preprint arXiv:1907.06333*, 2019.

[150] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, *et al.*, "The INTERSPEECH 2012 speaker trait challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA: ISCA, 2012.
[151] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," in *International Workshop on Affective Interactions*, Springer, 1999, pp. 150–165.
[152] H. Ritschel, I. Aslan, S. Mertes, A. Seiderer, and E. André, "Personalized synthesis of intentional and emotional non-verbal sounds for social robots," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, Cambridge, United Kingdom, 2019, pp. 1–7.
[153] A. Baird, S. Amiriparian, and B. Schuller, "Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation," in *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, Kuala Lumpur, Malaysia, 2019, pp. 1–5.
[154] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proceeidngs of the International Conference on Multimedia and Expo (ICME)*, Shenzhen, China: IEEE, 2021, pp. 1–6.
[155] A. Ando, R. Masumura, H. Sato, T. Moriya, T. Ashihara, Y. Ijima, and T. Toda, "Speech emotion recognition based on listener adaptive models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, ONT, Canada, 2021, pp. 6274–6278.
[156] S. Amiriparian, J. Han, M. Schmitt, A. Baird, A. Mallol-Ragolta, M. Milling, M. Gerczuk, and B. Schuller, "Synchronization in interpersonal speech," *Frontiers in Robotics and AI*, vol. 6, p. 116, 2019.
[157] 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 Review*, vol. 13, no. 1, pp. 44–50, 2021.
[158] H. Ritschel, T. Baur, and E. André, "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon, Portugal: IEEE, 2017, pp. 378–384.
[159] 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 *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, Montreal, QC, Canada, 2021, pp. 613–618.
[160] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Training socially engaging robots: Modeling backchannel behaviors with batch reinforcement learning," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–14, 2022.

Andreas Triantafyllopoulos obtained his Diploma in ECE from the University of Patras, Greece, in 2017. He is now working as a research assistant researcher at the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, where he has been pursuing his doctoral degree since 2018. His current focus is on deep learning methods for auditory intelligence and affective computing. He is a Student Member of the IEEE and its Signal Processing Society (SPS).

Björn W. Schuller (M'06 – SM'15 – F'18) received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professor all in EE/IT from TUM in Munich/Germany. He is Full Professor of Artificial Intelligence and the Head of GLAM at Imperial College London/UK, Full Professor and Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, co-founding CEO and current CSO of audEERING. He is a Fellow of the IEEE and Golden Core Awardee of the IEEE Computer Society, Fellow of the BCS, Fellow of the ISCA, President-Emeritus of the AAAC, and Senior Member of the ACM. He (co-)authored 1,200+ publications (45k+ citations, h-index=100+), is Field Chief Editor of Frontiers in Digital Health and was Editor in Chief of the IEEE Transactions on Affective Computing amongst manifold further commitments and service to the community.

Gökçe İymen received her B. S. degree in Industrial Engineering from Middle East Technical University, Turkey in 2019. She is currently studying for a M. S. degree in Data Science at Koç University, Turkey, where her research focuses on applications of deep learning for audio generation, especially adding affect to speech.

Metin Sezgin graduated summa cum laude with Honors from Syracuse University in 1999. He completed his MS in the Artificial Intelligence Laboratory at Massachusetts Institute of Technology in 2001. He received his PhD in 2006 from Massachusetts Institute of Technology. Dr. Sezgin is currently an Associate Professor in the College of Engineering at Koç University, Istanbul. His research interests include intelligent human-computer interfaces, multimodal sensor fusion, and HCI applications of machine learning. He has held visiting posts at Harvard University and Yale University. His research has been supported by international and national grants including grants from the European Research Council, and Turk Telekom. He is a recipient of the Career Award of the Scientific and Technological Research Council of Turkey.

Xiangheng He is a PhD candidate with GLAM – the Group on Language, Audio, & Music, Imperial College London, London, UK. She is currently also a research assistant with University of Augsburg, Germany. She received her Master degree from Southeast University, China, in 2020. Her research focuses on affective computing and voice conversion.

Zijiang Yang (Student Member, IEEE) received his Master degree in Information Technology from University of York, UK, in 2016. He is currently a research assistant and pursuing his Ph. D. degree with University of Augsburg, Germany. His research focuses on deep learning, affective computing and speech synthesis.

Panagiotis Tzirakis earned his Ph. D. with the Intelligent Behaviour Understanding Group (iBUG) at Imperial College London, 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 3D facial motion synthesis, multi-channel speech enhancement, the detection of Gibbon calls, and emotion recognition from audio and video.

Shuo Liu received his M. Sc. degree in electric engineering and information technology from Technical University of Darmstadt (TUD), Darmstadt, Germany, in 2017. He is currently pursuing his doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg. His current research interests include deep learning and machine learning algorithms for speech and audio processing, affective computing, and health-related applications.

Silvan Mertes is a PhD candidate at the Chair of Human-Centered Artificial Intelligence at Augsburg University in Germany, where he also received his M. Sc. degree in Computer Science in 2019. 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 is a full professor of Computer Science and Founding Chair of Human-Centered Artificial Intelligence at Augsburg University in Germany, and co-speaker of the Bavarian Research Association ForDigitHealth. She has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing and social signal processing. Her work has won many awards including the Gottfried Wilhelm Leibniz Prize, the most important research funding award in Germany, and she is a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina and the CHI Academy. In 2013, she was awarded a EurAI fellowship (European Association for Artificial Intelligence). Most recently, she was named one of the 10 most influential figures in the history of AI in Germany by National Society for Informatics (GI). Since 2019, she is serving as the Editor-in-Chief of IEEE Transactions on Affective Computing.

Ruibo Fu is an assistant professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He obtained his B. E. from Beijing University of Aeronautics and Astronautics in 2015 and Ph. D. from the Institute of Automation, Chinese Academy of Sciences in 2020. His research interests lie with speech synthesis and transfer learning. He has published more than 10 papers in international conferences and journals such as ICASSP and INTERSPEECH and has won the best paper award twice in NCMMSC 2017 and 2019. He won the first prize in personalised 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 ICASSP2021 Multi-Speaker Multi-Style Voice Cloning Challenge (M2VoC) Challenge.

Jianhua Tao received his PhD degree from Tsinghua University in 2001, and got his Ms from Nanjing University in 1996. He is currently a Professor in NLPR, Institute of Automation, Chinese Academy of Sciences. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing and pattern recognition. He has published more than eighty papers on major journals and proceedings including IEEE Trans. on ASLP, and received several awards from the fields' important conferences, such as Eurospeech, or NCMMSC. He serves as the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, or NCMMSC. He also serves as the steering committee member for IEEE Transactions on Affective Computing, associate editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions, and Deputy Editor-in-chief for Chinese Journal of Phonetics.