Music Theory-inspired Acoustic Representation for Speech Emotion Recognition

Xingfeng Li, Xiaohan Shi, Desheng Hu, Yongwei Li, Qingchen Zhang, *Senior Member, IEEE*, Zhengxia Wang, Masashi Unoki, *Member, IEEE*, Masato Akagi, *Life Member, IEEE*

***Abstract*—This research presents a music theory-inspired acoustic representation (hereafter, MTAR) to address improved speech emotion recognition. The recognition of emotion in speech and music is developed in parallel, yet a relatively limited understanding of MTAR for interpreting speech emotions is involved. In the present study, we use music theory to study representative acoustics associated with emotion in speech from vocal emotion expressions and auditory emotion perception domains. In experiments assessing the role and effectiveness of the proposed representation in classifying discrete emotion categories and predicting continuous emotion dimensions, it shows promising performance compared with extensively used features for emotion recognition based on the spectrogram, Mel-spectrogram, Mel-frequency cepstral coefficients, VGGish, and the large baseline feature sets of the INTERSPEECH challenges. This proposal opens up a novel research avenue in developing a computational acoustic representation of speech emotion via music theory.**

***Index Terms*—Affective computing, speech emotion recognition, acoustic representation, music theory and speech analysis.**

I. Introduction

EMOTION is a psychological state brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure [1–3]. It can function to communicate positively important information to individuals in significant external events, such as values and ethics [4]. Also, it is sometimes internally regarded as part of a mental illness and thus possibly of negative value, for instance, anxiety or depression [5, 6]. There is widespread evidence supporting that emotion is of vital importance for social competence, behavior, health, and well-being [7–11]. Since the early 1980s, many different disciplines have produced studies on emotions spanning psychology, medicine, history, sociology of emotions, and computer science [12–19]. As an interdisciplinary field spanning the aforementioned disciplines, affective computing aims to give machines emotional intelligence, including simulating empathy, and has been receiving increasing interest [18–21]. One of motivation for the research is the ability to recognize the emotional state of humans and adapt machines' behaviors to them by giving an appropriate response to those emotions like a human conversational partner would [18, 19]. The machines recognize emotional information largely through two different kinds of perception cues, i.e., visual (facial expression, body posture, gesture, etc.) and auditory patterns (speech, singing, vocal bursts, etc.) [1, 19, 22–26]. Other relevant neurophysiological cues of emotion include skin temperature, functional magnetic resonance imaging, electrocardiography, electromyography, and electroencephalography [27–29]. Research on emotion recognition can advance many applications, like distance education [30], social robots [31], video games [32], affective mirrors [33], and many others [34].

Despite the substantial advances in this area, speech emotion recognition (SER) still faces several challenges [35–37]. One of which is the design of outstanding features that best reflect the emotional content and should be robust against other properties of speakers, like identity, genders, ages, etc. [26, 38, 39]. Researchers since the early 2000s have struggled to understand the principles underlying voice production on the phonatory and articulatory levels and thereby measure the emotion differentiating parameters [40–42]. Most of the established acoustics related to subglottal pressure, transglottal airflow, and vocal fold vibration have been empirically documented, including speech rate, fundamental frequency (F0), formant frequencies, intensity, etc. [43–45]. Also, many attempts have been made to replicate a human auditory perception of emotion to capture acoustic characteristics [46–48]. The Mel-spectrogram and Mel-frequency cepstral coefficients (MFCCs), for example, have been shown to be important cues to affective speech content and have been successfully applied in this field [19, 36, 46]. Other examples of relevant auditory-inspired acoustics include long-term modulation spectral features [49, 50]. In addition, numerous studies have commonly combined different grouping features to represent emotional spoken messages, such as the predefined large-scale, brute-forced acoustic feature sets offered by the openSMILE toolkit [51–58]. While deep learning started in this field in the early 2010s, a widespread trend exists to explore the neural representation of SER problems [59, 60]. For example, VGGish, a pre-trained convolutional neural network (CNN) from Google [61], has been investigated and confirmed by [62], as well as [63] and [64]. Even this profusion of the aforementioned acoustics enables each study to capture many emotional characteristics comprehensively and reliably. However, it comes at the cost of severe difficulties in comparing results across works, and thus endangering the accumulation of empirical evidence [41, 57, 65]. The main reason may be attributed to the fact that only partial overlap acoustic features are used in different studies [40, 65, 66]. To overcome the aforementioned difficulties, finding and designing ideal acoustic parameters is crucial.

This contribution goes one step beyond current emotion

recognition algorithms and proposes a method for studying computational acoustics of speech emotion on the basis of music theory. Historically and today, vocal and music expression and perception have dominated the nonverbal communication of emotion [67–71]. A series of prior studies have focused on the overlapping neurophysiological, cognitive, and perceptual processes commonly found in music and speech [72–74]. Juslin and Laukka [40], in a comprehensive review of earlier research on the vocal expression of emotion in speech and music, listed 104 studies of vocal expression and 41 studies of music performance. They suggested that the acoustic cues that convey emotion in speech are similar to those in music. For example, both angry speech and music are characterized by fast speech, high volume, and high pitch levels. These facts motivated broader researchers to argue that investigation into the recognition of emotion in speech prosody may provide knowledge of emotion recognition in music [74–77]. However, it is theorized that the expression of emotion in speech and music is related evolutionarily [40, 78]. Also, the recognition of emotion in both stimuli has been proved to be developed in parallel [74]. Comparatively, however, limited research efforts have been mirrored in constructing a more powerful music theory-inspired acoustic representation (hereafter, MTAR) for SER. Nevertheless, music has showed and highlighted an advanced capability to represent, express and elicit emotions [79]. For example, in [80], Trehub et al. reported that live maternal singing has more sustained effects on infant arousal than does live maternal speech. Sauter et al. also confirmed that although children have difficulty judging emotions from speech, children as young as five years old are proficient at judging emotions from nonverbal vocalizations, affect bursts, and music [81]. Other research further solidifies the link between music and speech and supports the fact that both children's and adults' training in music leads to enhanced sensitivity to the emotional message delivered by human speech [74, 82]. As such, it has thus been hypothesized that investigating acoustic representation in interpreting emotion from music could benefit and advance the recognition of emotion in speech.

In line with these findings, this article recommends a music theory-inspired acoustic analysis of emotional speech from two domains of vocal emotion expression and auditory emotion perception to reach this understanding. The proposed features are based on music notation analysis of the fundamental frequency and representation of Mel filters, thus capturing the emotional speech signal's articulation and auditory properties. These features are evaluated in the SER tasks dominated by two well-known emotion theories: (1) classification of discrete emotions under the categorical framework, which characterizes speech emotions using categorical descriptors; and (2) estimation of continuous emotions (e.g., valence, arousal, and dominance) under the dimensional framework that describes speech emotions as points in an emotion space. To our knowledge, there is no previous SER attempt at applying music notation content into the F0 and Mel-filtering for the purpose of acoustic computation. We address this problem in this article.

The remainder of this article is structured as follows: Section 2 provides a brief overview of music notation content in music theory and structure, which are the basis of the recommendation proposed in this article; Section 3 details the description for generating the representation of music theory-inspired SER acoustics; Section 4 evaluates the proposed feature set on the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [83] and extensively compared it with ten well-known affective acoustic references in the research competitions in the field. The openSMILE toolkit offers six predefined feature sets broadly used in the INTERSPEECH Challenges on Emotion and Paralinguistics from 2009 to 2016. The Mel-spectrogram, MFCCs, spectrogram, and VGGish, commonly used in deep learning algorithms, serve as the remaining four baselines. Final remarks on the parameters recommended in this article and the categorical and dimensional SER performance relative to other established sets, as well as a discussion on the future research, are given in Section 5.

II. Related Motivate Work of MTAR

As apparent from the aforementioned review, our research is one of the first few contributions that provides new insight into SER in designing computational acoustics of emotion in a speech via investigating the processing of emotion evolution in music. This section provides a backdrop that explains the motivations for the current approach. This section focuses on (i) the acoustic representation that signals music emotions and its approximate counterpart in emotional speech; and (ii) the process that modulates the acoustic properties of music and speech for expressive emotion in parallel.

The first point for music and speech emotion relates to specifying approximate acoustic cues of emotion in both channels. Throughout history and across cultures [84–88], humans have created music that relies on a system of pitch intervals [89, 90]. This system enables for the perception and production of consonance and dissonance stimuli [84, 91]. In Western music, there exist twelve intervals [92]. Each interval has a unique feeling associated with it, mainly dominated by "happy, energetic, positive" or "sad, sentimental, negative" [93]. Further solidifying the link between interval and music emotion, singing lessons emphasize controlled use of the voice to produce expressive performances by a sequence of discrete pitches [94]. Also, even 5- to 8-year-olds proactively modulate pitch values in their performance while conveying singing emotion [95, 96]. Arguably, the ubiquitous role of intervals in music emotion is particularly striking and has thus been of considerable interest [90, 91, 97].

As Patel depicted [72], the same neural resources are involved in processing F0 contours in speech and pitch contours in music. Fujisawa et al. have previously studied the relationship between emotion perception and F0 on the basis of the music intervals, suggesting that interval structure is a fruitful means for determining the emotional valence of speech [98]. Yang and Lugger have also applied the harmony perception known from music to improve SER performance [99]. The findings confirmed that interval-inspired features associated with F0 counters are important cues to affective speech content. They further introduce a single dissonance

parameter, called DIS, to summarize the consonance and dissonance effects of all occurring pitch intervals. However, Lerdahl and Jackendoff noted that both music and language phrase structures have a hierarchical organization in parallel [100]. Unfortunately, the aforementioned studies focused on the interval features of F0 counters on a frame-level only, omitting important phrasing information in music and speech structures.

This article discusses the music notation of interval-inspired computational acoustics of emotion differently from previous studies in several ways. First, we introduced a phrase structure of emotional intervals that formed a hierarchical organization in speech from the frame to word levels. Second, although it is well-known that the auditory system analyzes the intervals [101, 102], still, there are relatively few general rules on how to extract good interval features from this domain. We originally addressed this problem by adopting the interval content to represent auditory emotion perception.

The second point was to investigate the approximate process in music and speech that deals with the interval content to deliver an expressive emotion. On the one hand, an early notion of 'musilanguage,' given by Brown, effectively studied the evolution of music and language in parallel [103–105]. Brown suggested that music and speech are hierarchical and melodorhythmic structures derived from three sources [104]. The first is the acoustic properties of fundamental units such as pitch sets and intensity values in music and phonemes and phonological feet in speech. The second refers to a sequential arrangement of the aforementioned units in a given phrase following their combinatorial syntax. The third is the expressive mechanisms that modulate the basic acoustic properties of the phrase for emphasis, emotion, and intention. In addition, hearing scientists have further confirmed that humans do not directly "record" the acoustic wave in listening. Instead, parameters are abstracted from the sounds to form compressed representations of the acoustic event [106]. For example, a variety of phonological units (e.g., phonemes, syllables, words) have been postulated as significant elements in the human data structure for speech comprehension [107]. These findings indicated that we depend critically on relating our auditory input to the over-learned interval structures that characterize the emotion of music and speech hierarchically [85, 108, 109]. In other words, this hierarchical mechanism requires the computation of acoustic relations between basic events, such as frames, phonemes, syllables, or words in speech, and notes, segments, syllables, or phrases in music [90]. This fact motivated us to identify the process that modulates speech acoustics for expressive emotions using a hierarchical syntactic structure.

III. Music Theory-inspired Acoustic Representation

Figure 1 describes a process when an emotional message in the speech chain travels [110, 111] from the speaker's mind to the listener's mind. It consists of (1) an encoding process on vocal emotion expression, in which a speaker produces words and generates affective sound waves and transmits the waveform through the air medium, and (2) a decoding process on auditory emotion perception in a listener's auditory system that perceives how it was said.

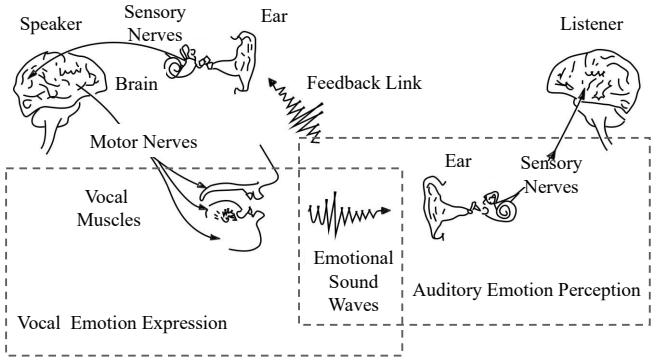

Fig. 1. Speech chain and related spoken emotion communication.

Most researchers believe that the two aforementioned processes derive much of the emotional content in an utterance [26, 42, 43, 48]. The following sections introduce the proposed attempt to design an 80-dimensional MTAR of emotional speech along with time span via investigating music theory contents, including a 41-dimensional MTAR in vocal emotion expression and a 39-dimensional MTAR in auditory emotion perception processes.

A. MTAR of Vocal Emotion Expression

Musical notation is the most human-readable and the oldest symbolic representation system that stores the structure of music events [112]. It could provide instruction on how to perform the music and thus dominates the expression of music events and their emotions [89, 113]. Note that both music and speech can express emotions and share similarities in their vocalization [40, 74, 79]. However, to our knowledge, no previous work is particularly comprehensive in studying such content of notation in the representation of vocal emotion expression to approach SER tasks. This article first addresses this content and studies the following notation of 41-dimensional MTAR, inspired by five music theory groups.

a) MIDI note related descriptors: A MIDI note is a digital representation of a musical note in the MIDI protocol. It can include parameters of velocity (how hard a key was struck), duration, and other musical performance data. Nowadays, advances in music emotion recognition (MER) have been evident by exploring MIDI note-related features [71, 114]. For instance, Panda et.al, computed 6 statistics: MIDImean, MIDIstd, MIDIskew, MIDIkurt, MIDImax, and MIDImin to capture the melody characteristics for MER [115]. Other examples of the promise for MER by exploring MIDI notes also include [116]. Most interestingly, the MIDI note number and pitch are formally related and can be converted. This contribution, therefore, proposed the following nine pitch-driven MIDI note related descriptors to approach SER.

- **Pitch**, i.e., F0 estimates from the *pitch* function in MATLAB_R2021a, specifying an *'SRH'* method [117] with a window size of 40 ms and an overlap of 30 ms.

- **MIDI note number**, mathematically, the MIDI note number, $m(t)$, is given by

$$m(t) = \lfloor | \, S \log_2 \left(F\left(t\right) / F_{ref} \right) + m_{ref} \mid + 0.5 \rfloor \quad (1)$$

where, $F(t)$ is an F0 value at the t frame, $S = 12$ semitones per octave are used in Western music [89, 118]. The MIDI standard proposes $F_{ref} = 440$ Hz for the standard pitch $A4$ and $m_{ref} = 69$ to count pitches [119]. We determine the $m(t)$ by rounding it to the nearest key.
- **MIDI notes and octaves**, two fundamental components to specify a MIDI note number with a particular musical pitch. We accordingly introduce two numeric embedding vectors $\eta\left(t\right)$ and $o\left(t\right)$ to represent the aforementioned components. Table I shows the mapping between MIDI note numbers, notes, and octaves using the scientific pitch notation (SPN).
- **MIDI relative note number**, $rel_{m(t)}$, the ratio of the current $m(t)$ to the totaling MIDI note numbers using the SPN.
- **MIDI note number run-length encoding (RLE)**, $rle_{m(t)}$, summarizes the counts of the current $m(t)$ that occurs in the whole emotional speech. For instance, let $M_{example} = [69, 69, 70, 70, 72]$ be a MIDI note number sequence in an emotional speech. The resultant $rle_{m(t)}$ sequence is $[2, 2, 2, 2, 1]$.
- **Relative counts of the current $m(t)$ in the MIDI note number RLE**, $REL^{rle}_{m(t)}$, ratio of the counts of $m(t)$ that occurs over the current moment to that in the whole emotional speech. The $REL^{rle}_{m(t)}$ is given as $\left[\frac{1}{2}, \frac{2}{2}, \frac{1}{2}, \frac{2}{2}, \frac{1}{1}\right]$, which refers to $M_{example}$ and $rle_{m(t)}$ sequences.
- **MIDI note number ascending and descending,** $asc(t)$, measures a direction in emotional pitch modulation, which is also important to signal other dimensions of musical notation, such as intervals, where C to E is a major 3rd ascending while E to C is a major 3rd descending.

$$asc(t) = \begin{cases} 1, & m(t) < m(t+1) \\ 0, & m(t)\text{=}m(t+1) \text{ or } t\text{=}T \\ -1, & \text{otherwise.} \end{cases}$$

- **MIDI note number range**, formulated by quantiles of a sequence $M_t = [m\left(1\right), m\left(2\right), m\left(t\right), ..., m\left(T\right)]$ corresponds to the distribution between the lowest to highest pitch uttered in a speech signal and significantly varied among speakers, genders, and emotional states [43, 65, 66]. This contribution investigates a numeric vector $q(t) \in \{1, 2, 3, 4, 5, 6\}$, which is given by,

$$q(t) = \begin{cases} 1, & m(t) < Q(M_t, .025) \\ 2, & m\left(t\right) \in \left(Q\left(M_t, .025\right), Q\left(M_t, .25\right)\right] \\ 3, & m\left(t\right) \in \left(Q\left(M_t, .25\right), Q\left(M_t, .5\right)\right] \\ 4, & m\left(t\right) \in \left(Q\left(M_t, .5\right), Q\left(M_t, .75\right)\right] \\ 5, & m\left(t\right) \in \left(Q\left(M_t, .75\right), Q\left(M_t, .95\right)\right] \\ 6, & \text{otherwise.} \end{cases}$$

TABLE I
Q-TABLE FOR CONVERTING MIDI NOTE NUMBER ($m(t)$) INTO NOTES ($\eta\left(t\right)$), OCTAVES ($o\left(t\right)$), AND VICE VERSA

	$\eta\left(t\right)$	1	2	3	4	5	6	7	8	9	10	11	12
$o\left(t\right)$	Note / Octave	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
1	-1	0	1	2	3	4	5	6	7	8	9	10	11
2	0	12	13	14	15	16	17	18	19	20	21	22	23
3	1	24	25	26	27	28	29	30	31	32	33	34	35
4	2	36	37	38	39	40	41	42	43	44	45	46	47
5	3	48	49	50	52	52	53	54	55	56	57	58	59
6	4	60	61	62	63	64	65	66	67	68	69	70	71
7	5	72	73	74	75	76	77	78	79	80	81	82	83
8	6	84	85	86	87	88	89	90	91	92	93	94	95
9	7	96	97	98	99	100	101	102	103	104	105	106	107
10	8	108	109	110	111	112	113	114	115	116	117	118	119
11	9	120	121	122	123	124	125	126	127	128	129	130	131
12	10	132	133	134	135	136	137	138	139	140	141	142	143

Mathematically, given the MIDI note number, $m(t)$, the embedding note($\eta\left(t\right)$), is given by ($m(t)$ % 12)+1; the embedding octave($o\left(t\right)$), is given by 1+floor($m(t)$/12), using SPN.

b) Music dynamics related descriptors: dynamics are one of the expressive elements in music theory, helping musicians sustain variety and interest in a musical performance and communicate a particular emotional state or feeling [120, 121]. The dynamics of a piece refer to the variation in loudness between notes and phrases. Similarly, we study this dynamic-inspired notation in a speech by exploring three intensity-related descriptors in terms of the nonlinear Teager energy operator (TEO) [122, 123] at frame levels. TEO allows for a dynamic assessment of energy based on the signal's amplitude and instantaneous frequency content, making it highly sensitive to subtle emotional changes in speech.

- **Dynamic markings**, $\zeta(t)$, refers to the summation of the TEO within one frame t and is given as follows:

$$\zeta(t) = \sum_{n=1}^{\iota} \left(\psi_{x(n)}\right) \quad (2)$$

where ι is the length of total sampled speech signal, $x(n)$, in one frame t of 10 ms that is consistent with the specification while estimating the F0 counter. Moreover, the TEO introduced by Kaiser [122, 123] is defined as:

$$\psi_{x(n)} = x\left(n\right)^2 - x\left(n+1\right) x\left(n-1\right) \quad (3)$$

- **Dynamic variations**, i.e., the a) mean ($\zeta_{avg}(t)$) and b) standard deviation ($\zeta_{std}(t)$) in the frame-level TEO values between a continuous MIDI note number.

$$\zeta_{avg}(t1:t2) = \left(\frac{1}{L_\zeta}\right) \times \sum_{t1}^{t2} \zeta\left(t\right) \quad (4)$$

$$\zeta_{std}\left(t1:t2\right) = \sqrt{\left(\frac{1}{L_\zeta}\right) \times \sum_{t'=t1}^{t2} \left(\zeta\left(t'\right) - \zeta_{avg}\left(t0\right)\right)^2} \quad (5)$$

where, $L_\zeta = t2 - t1 + 1$, $t0 = \forall t \in \{t1, ..., t2\}$ and $t1$ and $t2$ are the indexes of beginning and ending frames, respectively, of a continuous MIDI note number in the M_t sequence.

c) Music main interval-related descriptors: An interval is a difference in pitch between two sounds [124]. It could be described as horizontal or melodic if it refers to successively sounding tones, such as two adjacent pitches in a melody, and vertical or harmonic if it pertains to simultaneously sounding tones, such as in a chord [125]. Intervals can be categorized into consonance and dissonance groups [93]. Within the Western tradition, several listeners associate consonance with sweetness, pleasantness, and acceptability, and dissonance with harshness, unpleasantness, or unacceptability, although there is broad acknowledgement that this depends also on familiarity and musical expertise [126, 127]. Due to the pitch's intrinsic temporal nature in speech, we thus study the following interval-related descriptors in a melodic manner.

- **Semitones, interval orders, and types**, in Western music theory, different intervals are mostly varied in spanning semitones in addition to the intervals of $augmented\ 4th$ and $diminished\ 5th$ that both span six semitones. These intervals can be named in accordance with its order (also called a diatonic number) and type. For instance, a major third (or $M3$) is an interval name spanning four semitones, in which the term major (M) describes the type of the interval, and third (3) indicates its order. This contribution studies the semitones, $\varsigma(t)$, on the basis of MIDI note number, and refers to Table II for converting semitones into interval orders ($\sigma(t)$) and types ($\tau(t)$).

$$\varsigma\left(t\right) = \mid \left(m\left(t+1\right) - m\left(t\right)\right) \mid \%12 \tag{6}$$

in particular,

$$\varsigma\left(t\right) = \begin{cases} 0, & \text{if } m\left(t+1\right) = m\left(t\right) \mid t = T \\ 12, & \text{if } m\left(t+1\right)/m\left(t\right) \in \{0.5, 2\} \\ \varsigma\left(t\right), & \text{otherwise.} \end{cases}$$

- **Main intervals RLE**, $rle_{\varsigma(t)}$, summarizes the counts of the current $\varsigma(t)$ that occurs in the whole emotional speech. For example, the main interval sequence, $I_{example}$, is defined on the basis of $M_{example}$, and being $[0, 1, 0, 2, 0]$. The $rle_{\varsigma(t)}$ sequence is thus given as $[3, 1, 3, 1, 3]$.
- **Relative counts of the current $\varsigma(t)$ in the main interval RLE**, $REL^{rle}_{\varsigma(t)}$, the ratio of the counts of $\varsigma(t)$ that occurs over the current moment to that occurs in the whole emotional speech. The $REL^{rle}_{\varsigma(t)}$ is given as $\left[\frac{1}{3}, \frac{1}{1}, \frac{2}{3}, \frac{1}{1}, \frac{3}{3}\right]$ refers to $I_{example}$ and $rle_{\varsigma(t)}$ sequences.

d) Microtonal music-related descriptors: the term *'microtonal music'* usually refers to music containing very small intervals that can be expressed as a rational ratio of pitch frequencies [128]. Still, it can include any tuning that differs from Western twelve-tone equal temperament [125, 129, 130]. This contribution introduces nine microtonal music-related descriptors that are dominated by the F0 sequence, *F(t)*.

- **Cent**, is a unit of measure for the ratio between two frequencies [131]. Typically, it is used to express small intervals and compare the sizes of comparable intervals in different tuning systems [132, 133]. Formally, let *F(t)* and *F(t+1)* be two adjacent frames throughout the F0 of an emotional speech. A sequence of cents, $\chi\left(t\right)$, can be given as follows:

$$\chi\left(t\right) = 1200 \times \log_2 f^{ratio} \tag{7}$$

where,

$$f^{ratio}\left(t\right) = \frac{\max\left(F(t), F(t+1)\right)}{\min\left(F(t), F(t+1)\right)} \tag{8}$$

For instance, let *F(t)* $= [15, 16, 14, 13, 15]$ being the F0 sequence of an emotional speech. The resultant $\chi\left(t\right)$ sequence is $[111.73, 231.17, 128.30, 247.74, 0]$.

- **Monzo**, a *p-limit* rational number f^{ratio} can by definition be factored into primes of sizes less than or equal to p, giving $f^{ratio} = 2^{e_2}3^{e_3}5^{e_5}...p^{e_p}$, where the exponents are integers (positive, negative, or zero). This is often written in a *ket* vector notation as $\mid e_2 e_3 e_5 ... e_p\rangle$. It enables expressing directly how any "composite" interval is represented in terms of those simpler prime intervals. This research studied the interval's prime factorization, up to the first three primes. For example, the 5-limit interval $f^{ratio}\left(t\right)$=16/15 factors as $2^4 3^{-1} 5^{-1}$, so it has a three-dimensional monzo representation of $Z\left(t\right)$=[4 -1 -1$\rangle$.
- **Benedetti height and *no-2s* Benedetti height**, are known as a measurement of product complexity [134, 135]. The Benedetti height ($\beta_{height}\left(t\right)$) of a positive rational number f^{ratio} reduced to the lowest terms is given by the product of the numerator and denominator as:

$$\beta_{height}\left(t\right) = F(t) \times F(t+1) \tag{9}$$

where *F(t)* and *F(t+1)* have no common factors. In addition, the *no-2s* Benedetti height, $\beta^{no-2s}_{height}\left(t\right)$, is defined by (9), but removing all factors of *2* before computing the height. For example, when two adjacent F0 frames, *F(t)* and *F(t+1)*, give an interval, $f^{ratio}\left(t\right)$=16/15, the $\beta_{height}\left(t\right)$ is then equal to 16×15 and $\beta^{no-2s}_{height}\left(t\right)$=15.
- **Tenney height and *no-2s* Tenney height**, are the logarithm base two of the Benedetti height and *no-2s* Benedetti height, respectively. They can be thought of as measures of a ratio's harmonic complexity or dissonance. Mathematically, the Tenney height is given by

$$\Gamma_{height}\left(t\right) = \log_2 \beta_{height}\left(t\right) \tag{10}$$

TABLE II
Q-TABLE FOR CONVERTING SEMITONES ($\varsigma(t)$) INTO INTERVAL ORDERS ($\sigma(t)$) AND TYPES ($\tau(t)$)

Semitones	Music Theory		Music-Inspired Notation	
$\varsigma(t)$	Order	Type	Order($\sigma(t)$)	Type($\tau(t)$)
0	First	Perfect	1	0
1	Second	Minor	2	-1
2	Second	Major	2	1
3	Third	Minor	3	-1
4	Third	Major	3	1
5	Fourth	Perfect	4	0
6	Fifth	Diminished	5	-2
7	Fifth	Perfect	5	0
8	Sixth	Minor	6	-1
9	Sixth	Major	6	1
10	Seventh	Minor	7	-1
11	Seventh	Major	7	1

and the *no-2s* Tenney height is obtained by:

$$\Gamma_{height}^{no-2s}(t) = \log_2 \beta_{height}^{no-2s}(t) \tag{11}$$

- **Degree**, $\delta(t)$, is a measure that associates with the step that represents how to view pitch intervals in an equal-step tuning such as an equal division of the octave (EDO). Formally, let $\lambda(t)$ be a step sequence whose values are determined by the *patent val* vector, ν=⟨12 19 28], in a 12EDO scale and the monzo sequence, $Z(t)$=[e$_2$ e_3 e_5⟩. The $\delta(t)$ can be given as follows:

$$\delta(t) = \begin{cases} \lambda(t) - 1, & \text{if } \lambda(t) < 0 \\ \lambda(t) + 1, & \text{otherwise.} \end{cases}$$

where,

$$\lambda(t) = \langle \nu \mid Z(t) \rangle = 12 \times e_2 + 19 \times e_3 + 28 \times e_5 \tag{12}$$

e) Music syntactic structure related descriptors: given that humans depend critically on relating auditory input to the melodorhythmic structure that conveys an emotional expression of speech and music hierarchically between basic events [90, 104, 106, 107], this research designs a simplified group of statistical descriptors to represent the syntactic structure. As shown in Figure 2, the descriptors include the following 14 items:

- $\mathbf{Ut_f}$, the number of frames in the current utterance.
- $\mathbf{REL^f_{Ut}}$, the relative-position of the current frame in the current utterance.
- $\mathbf{Ut_p}$, the number of phonemes in the current utterance.
- $\mathbf{REL^p_{Ut}}$, the relative-position of the current phoneme in the current utterance.
- $\mathbf{Ut_s}$, the number of syllables in the current utterance.
- $\mathbf{REL^s_{Ut}}$, the relative-position of the current syllable in the current utterance.
- $\mathbf{Ut_w}$, the number of words in the current utterance.
- $\mathbf{REL^w_{Ut}}$, the relative-position of the current word in the current utterance.
- $\mathbf{Ph_f}$, the number of frames in the current phoneme.
- $\mathbf{REL^f_{Ph}}$, the relative-position of the current frame in the current phoneme.
- $\mathbf{Sy_f}$, the number of frames in the current syllable.
- $\mathbf{REL^f_{Sy}}$, the relative-position of the current frame in the current syllable.
- $\mathbf{Wd_f}$, the number of frames in the current word.
- $\mathbf{REL^f_{Wd}}$, the relative-position of the current frame in the current word.

B. MTAR of Auditory Emotion Perception

In this section, we describe the technical details for the proposed MTAR of auditory emotion perception; see Figure 3 for an overview. As for the front-end transform, the emotional sound waves are first aurally filtered using 40 Mel filters, specifying a Hamming window with a *40*-ms window size and *10*-ms shifts. It is a well-known fact that loudness is perceived in a logarithmic fashion [136, 137]. Therefore, the auditory emotion representation is given by a Log-Mel spectrogram (Mel_{log}).

$$Mel_{log} = \begin{bmatrix} \alpha_{1,1} & \cdots & \alpha_{1,t} & \cdots & \alpha_{1,T} \\ \alpha_{2,1} & \cdots & \alpha_{2,t} & \cdots & \alpha_{2,T} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{nBands,1} & \cdots & \alpha_{nBands,t} & \cdots & \alpha_{nBands,T} \end{bmatrix} \tag{13}$$

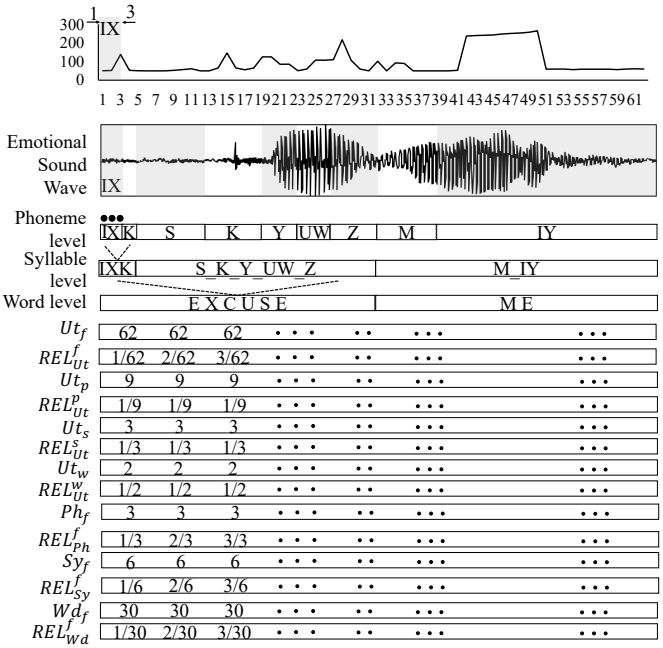

Fig. 2. Example of extracting statistical descriptors to study a syntactic structure in an emotional speech.

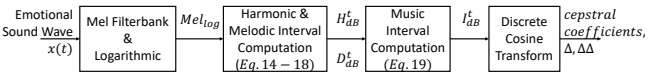

Fig. 3. Block diagram for deriving the MTAR of auditory emotion perception.

where, $nBands = 40$ is the number of Mel-bandpass filters.

We then proceed in three-stages to study the musical interval contents in Mel_{log} to obtain a music interval representation. First, the harmonic interval pertaining to two simultaneously sounding signal components at frequencies of f_1 and f_2 in each frame t is defined as:

$$H_{dB}^{t}(f_1, f_2) = 10 \lg \left((P_{(f_1,t)dB}^2 + P_{(f_2,t)dB}^2)/P_{ref}^2 \right) \tag{14}$$

and,

$$P_{(f,t)dB}^2 = 10^{(Mel_{log}^t(f)/10)} P_{ref}^2 \tag{15}$$

where,P_{ref}^2=$4 \times 10^{-10} N^2/m^4$ is a sound reference pressure and $\{(f_1, f_2) \mid f_1 \in \{1, ..., nBands\}, f_2 \in \{f_1, ..., nBands\}\}$. In particular, $\forall f_1 = f_2 : H_{dB}^t(f_1, f_2) = Mel_{log}^t(f1) = Mel_{log}^t(f2)$. Then, the melodic interval referring to successively paired sounding signal components in two adjacent frames is given by:

$$D_{dB}^t(f_1, f_2) = 10 \lg((10^{(D_{1,dB}^t(f_1,f_2)/10)} P_{ref}^2$$

$$+10^{(D_{2,dB}^t(f_1,f_2)/10)} P_{ref}^2)/P_{ref}^2) \tag{16}$$

where,

$$D_{1,dB}^t(f_1, f_2) = 10 \lg \left((P_{(f_1,t)dB}^2 + P_{(f_2,t+1)dB}^2)/P_{ref}^2 \right) \tag{17}$$

$$D_{2,dB}^t(f_1, f_2) = 10 \lg \left((P_{(f_1,t+1)dB}^2 + P_{(f_2,t)dB}^2)/P_{ref}^2 \right) \tag{18}$$

notably, $\forall f_1 = f_2 : D^t_{dB}(f_1, f_2) = D^t_{1,dB}(f_1, f_2) = D^t_{2,dB}(f_1, f_2)$. Next, the interval representation of an emotional sound wave is obtained:

$$I^t_{dB}(f_1, f_2) = 10\lg((10^{(H^t_{dB}(f_1,f_2)/10)}P^2_{ref} + 10^{(D^t_{dB}(f_1,f_2)/10)}P^2_{ref})/P^2_{ref}) \quad (19)$$

Finally, we calculate the cepstral coefficients of I^t_{dB} (with delta and delta-deltas) by a discrete cosine transform, resulting in a 39-dimensional MTAR of auditory emotion perception in a time-frequency scale.

IV. Baseline Evaluation

Most feature sets that have been used for SER can be divided into two categories: one-dimensional and two-dimensional acoustic feature sets. One-dimensional acoustic feature sets offer an excessive amount of brute-force parameters. Notable ones are widely given by the openSMILE toolkit, providing six acoustic baseline sets of the INTERSPEECH 2009 Emotion Challenge [51] (384 parameters), the INTERSPEECH 2010 Paralinguistic Challenge [52] (1582 parameters), the INTERSPEECH 2011 Speaker State Challenge [53] (4368 parameters), the INTERSPEECH 2012 Speaker Trait Challenge [54] (6125 parameters), the INTERSPEECH 2013 Computational Paralinguistics Challenge [138] (6373 parameters), which is also used for the INTERSPEECH 2014 [55] and 2016 Computational Paralinguistics Challenges [58], and the Geneva Minimalistic Acoustic Parameter Set [57] (GeMAPS, 88 parameters). On the other hand, two-dimensional acoustic feature sets that widely explore Mel-spectrogram (including delta and delta-delta, 120 dimensions), MFCCs (including delta and delta-delta, 39 dimensions), spectrogram (129 dimensions), and VGGish (128 dimensions) successfully applied to deep learning algorithms for SER, serving as another four baselines for comparison [19, 26, 48, 59, 60]. In this context, we showed the relevance of our proposed MTAR by comparing it to ten aforementioned well-known affective acoustic references by considering both one-dimensional and two-dimensional acoustic feature sets. Experiments are conducted to recognize discrete emotional categories and estimate continuous emotional dimensions via leave-one-speaker-out (LOSO) cross-validation (CV), where we use all utterances of each speaker once as the test set and each time use the utterances of their pair as the validation set.

A. Emotional Speech Data

We use the IEMOCAP corpus for this study for the following four reasons. First, this corpus is superior to alternative ones traditionally concerned with a scripted emotional speech. It further focused on spontaneous speech that contributes to pushing the application-oriented research on authentic emotions. Second, this corpus differs from conventional emotion annotations only by a limited number of discrete categories; it additionally presents a way to annotate emotional states in a multi-dimensional emotion space spanned by valence, arousal, and dominance. This dimensional annotation scheme contributes to an important challenge in SER, i.e., recognizing emotions continuously by identifying their categories and degrees or intensities with specific emotional states along the time plane. Third, this corpus defines precise alignments of primary events in an emotional speech in terms of phonemes, syllables, and words. This information could readily relate human auditory input to the over-learned hierarchical structures that are the same in modulating expressive emotions via segments and syllables, and phrases in music. Thus, this corpus is well-suited and effective to evaluate and verify the proposed MTAR for SER. Fourth, this publicly available database is one of the most commonly-used databases used for emotion recognition, and thus facilitating comparisons with other works. The IEMOCAP corpus has 12 hours of speech data from ten subjects and is pre-segmented into shortcuts, resulting in a total of 10,039 utterances. It contains nine categorical and three-dimensional labels of emotion. Experiments in this article are two-fold to evaluate the performance of the proposed MTAR for SER. In the first stage, we studied categorical emotions with majority labeling, where at least two annotators agreed on the annotation labeling. We consider four categorical emotions consistent with experimental protocols used in many previous studies, involving neutral state, happiness, sadness, and anger. Moreover, we merged happiness and excitement into one category. In the second stage, we explore continuous dimensional emotions of valence, arousal, and dominance, which enables for an adequate description of emotions regarding gradual emotion transitions and changes in the intensity of emotion.

B. Experimental Protocol

We begin with IEMOCAP, which does not contain a standard train/dev/test split. A common formulation, which we also adopt here, is to use LOSO CV [19, 26, 48, 59, 60]. The categorical SER performance is first evaluated by the average recall over the four different emotional discrete labels, termed unweighted accuracy (UA). We next report the following concordance correlation coefficient (CCC) metrics to evaluate the three-dimensional SER performance, by measuring the agreement between the outputs of the model and the ground truth.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (20)$$

where x and y are the predictions and annotations, μ_x and μ_y are the mean values, σ_x^2 and σ_y^2 are their variances, and ρ is the correlation coefficient.

As computational models, three neural networks are experimented with for categorical and dimensional SER tasks. The first model is a recurrent neural network (RNN), which is sensitive to capturing the changes in emotional states within a specific time slice [139–141]. The gated recurrent unit (GRU) network is an enhanced version of RNN that has the advantage of handling the vanishing gradient problem [142]; we use BiGRU in our model. This BiGRU is set to two layers with 256 units and a dropout rate of 0.5. In the training process, the optimizer is set to the Adam optimizer with a learning rate of 0.0001. In particular, this study follows [143], which uses a global average pooling (GAP) layer instead of a fully-connected layer to obtain the SER results. The GAP layer is

superior in reducing the model's parameters and preventing the model from being overfitted. The second computational model is designed using the CNN. It consists of four convolutional layers. The first layer comprises 32 filters with a kernel size of 5x5 and stride 2. The remaining three layers have 64, 96, and 128 filters, respectively, with the same kernel size of 3x3 and stride 2. The activation function of each layer is the rectified linear unit (Relu), and a dropout rate of 0.2 is applied. In addition, we flatten the last convolutional layer along the time axis, and then the time average pooling (TAP) is applied to extract time information of deep features. The output of the TAP layer is next mapped for SER tasks by a fully-connected layer and softmax activation layer. The third model is a temporal convolutional network (TCN), a new family of CNNs for sequence modeling, promising improved SER performance [144, 145]. We replace the last convolutional layer of the designed CNN model with a TCN block. It includes three one-dimensional dilated convolutional layers, each of which consists of one-dimensional dilated convolutions, layer normalization, elu activation function, and dropout. A residual connection is applied to the input and output of the TCN block. As a common formulation, we increase d exponentially with the depth of the network, i.e., the first TCN block with dilation factors d= 1, 2, 4. The kernel size of one-dimensional dilated convolutions is 3x3 and the number of dimensions of the output channel is 128. A dropout probability of 0.2 is used. Most notably, this study uses the loss function of cross-entropy for categorical SER tasks and that of mean square error for dimensional SER tasks in all three aforementioned computational models.

C. Results

This study compares the results obtained with the proposed MTAR with the most widely used acoustic representations in the field of paralinguistics of the Mel-spectrogram, MFCCs, spectrogram, and VGGish [19, 26, 48, 59, 60], and the large state-of-the-art brute-forced parameter sets from the series of INTERSPEECH Challenges on Emotion in 2009 [51] (InterSp09), Age and Gender as well as level of interest in [52] (InterSp10), Speaker State in 2011 [53] (InterSp11), Speaker Trait in 2012 [54] (InterSp12), Computational Paralinguistics in 2013, 2014, and 2016 [55, 58, 138] (InterSp13), the Geneva Minimalistic Acoustic Parameter Set [57] (GeMAPS), and the state-of-the-art studies in SER literature.

Table III shows the summarized results obtained for categorical SER. In particular, the column labeled "PIA" indicates the percentage improvement of the UA obtained by using the proposed MTAR to baseline sets, which is calculated as:

$$PIA = \begin{cases} \frac{MTAR_t^M - Base_t^M}{Ground_t - Base_t^M}, & \text{if} M \subseteq \{BiGRU, CNN, TCN\} \\ \frac{MTAR_t^{TCN} - Base_t^M}{Ground_t - Base_t^M}, & \text{otherwise} \end{cases}$$

where t represents the values of recognition rate (RR) or CCC, $Ground_{RR}$=100, and $Ground_{CCC}$=1.

As can be observed that classification results in terms of UA turn out to receive a notable gain from the proposed 80-dimensional MTAR in comparison with that of two individual feature subsets: a 41-dimensional MTAR in vocal emotion expression (VEE) and a 39-dimensional MTAR in auditory emotion perception (AEP), providing a PIA value ranging from 3.7% to 17.9%. Such results resonate with the fact that exploring emotional acoustic combinations from VEE and AEP domains is the promise for an improved SER. Besides, the proposed MTAR consistently outperformed four common formulations of two-dimensional acoustic representations in SER tasks: spectrogram, Mel-spectrogram, MFCCs, and VGGish by means of BiGRU-, CNN-, and TCN-based computational models. It yields a notable PIA value ranging from 3.4% to 12.8%. This finding in turn indicates that the effective implementation of acoustic representation inspired by music theory indeed contributes to SER performance. It achieves the best accuracy in most emotions and in overall performance, reaching up to a 63.2% average RR using the TCN-based model. As an aside, it is worth noting that compu-

TABLE III
CLASSIFICATION RESULTS OF CATEGORICAL EMOTIONS BY DIFFERENT ACOUSTIC REPRESENTATIONS AND THE STATE-OF-THE-ART WORKS.

Methods	Features	Neu.	Hap.	Ang.	Sad.	UA	PIA
BiGRU	**MTAR**	**61.3**	**53.0**	63.0	70.1	**61.9**	-
	VEE	44.7	40.4	**68.3**	61.2	53.6	17.9%
	AEP	57.8	50.5	44.7	**73.1**	56.5	12.4%
	Spectrogram	55.2	49.1	60.4	65.0	57.4	10.6%
	MelSpectrogram	56.8	52.6	57.6	**73.1**	60.0	4.8%
	MFCCs	59.0	46.8	50.5	70.8	56.8	11.8%
	VGGish	57.2	52.3	61.0	71.0	60.4	3.8%
CNN	**MTAR**	55.0	49.0	**70.1**	**75.7**	**62.5**	-
	VEE	50.7	40.8	60.4	69.0	55.2	16.3%
	AEP	54.9	46.5	67.3	73.7	60.6	4.8%
	Spectrogram	46.9	49.0	60.7	71.3	57.0	12.8%
	MelSpectrogram	**58.7**	49.3	66.6	68.9	60.9	4.1%
	MFCCs	51.0	**52.0**	66.4	70.2	59.9	6.5%
	VGGish	56.9	44.2	59.7	68.7	57.4	12%
TCN	MTAR	**58.6**	47.9	71.2	75.1	**63.2**	-
	VEE	49.0	42.4	65.3	68.9	56.4	15.6%
	AEP	48.5	51.1	**73.6**	74.1	61.8	3.7%
	Spectrogram	57.7	46.3	64.9	72.6	60.4	7.1%
	MelSpectrogram	52.3	**54.1**	67.2	74.0	61.9	3.4%
	MFCCs	47.7	50.7	69.7	**76.2**	61.1	5.4%
	VGGish	56.4	45.4	63.7	66.9	58.1	12.2%
SVM	InterSp09	51.1	43.7	66.3	66.9	57.0	14.4%
	InterSp10	55.3	47.2	62.6	55.3	55.1	18.0%
	InterSp11	52.9	47.5	60.0	53.0	53.3	21.2%
	InterSp12	53.0	49.0	60.5	52.1	53.7	20.5%
	InterSp13	55.4	50.2	61.3	52.7	54.9	18.4%
	GeMAPs	49.0	41.0	62.5	71.8	57.3	13.8%
DNN	InterSp09	55.3	53.8	60.9	68.4	59.6	8.9%
	InterSp10	56.5	54.7	71.3	65.3	61.9	3.4%
	InterSp11	55.0	57.2	65.8	66.1	61.0	5.6%
	InterSp12	50.2	54.4	70.2	69.6	61.1	5.4%
	InterSp13	55.5	50.8	69.7	69.3	61.3	4.9%
	GeMAPs	58.0	48.9	66.1	69.2	60.6	6.6%
ACNN + AE [146]	MelSpectrogram	51.8	52.5	66.3	67.6	59.6	8.9%
Self-Attn + LSTM [147]	Raw audio	-	-	-	-	55.6	17.1%
CNN-ELM + STC [148]	Spectrogram	62.9	61.0	70.3	52.4	61.6	4.2%
eResNet [149]	MelSpectrogram	-	-	-	-	56.5	15.4%
MCED [150]	IS16_ComparE	-	-	-	-	61.4	4.7%
MEnAN [151]	MelSpectrogram	-	-	-	-	59.9	8.2%
CNN_SeqCap [152]	Spectrogram	-	-	-	-	56.9	14.6%

Neutral (Neu.), Happiness (Hap.), Anger (Ang.), and Sadness (Sad.)

tational models employing CNN and TCN architectures have mostly been shown to excel in recognizing negative emotional states, whereas those based on BiGRU exhibit superior performance in identifying neutral and positive emotional states. Such results suggest that negative emotions, such as anger and sadness, often display abrupt and intense fluctuations in emotional expression over time [153], which might be better captured by the short-term memory of CNN and TCN models. Whereas, neutral and positive emotions, such as happiness [154, 155], tend to exhibit more gradual and sustained changes in emotional expression, which are better captured by the long-term memory of BiGRU models. It has therefore demonstrated the promise of fused models for an improved SER. Furthermore, the large state-of-the-art brute-forced parameter sets from the series of INTERSPEECH Challenges on Emotions and Paralinguistics from 2009 to 2016 are also considered for comparison. It is clear from Table III that the proposed acoustic representation consistently outperforms the widely used openSMILE features. This might be due to the fact that the openSMILE features at a global level completely omit important temporal information in an emotional speech, which hampered the SER performance. In particular, differing from the above four two-dimensional acoustic feature sets and the proposed MTAR, the openSMILE feature sets are one-dimensional parameters. This study uses two popular recognizers to train an emotional speech model that can better handle the aforementioned brute-forced parameter sets. The first one chosen is the most widely-used static classifier in the field of paralinguistics: support vector machines (SVMs). In this study, the support vector classification is performed by the scikit-learn toolkit [156]. The linear kernel is used in categorical SER tasks, and 10^6 of maximum iteration. To reduce the data-unbalance influence, we select the "balance" parameter in the support vector classifier (SVC) model, and the other parameters are left as default. The second one is a deep neural network (DNN)-based model. It consists of three fully-connected layers, where the node parameters of each layer are 512, 256, and 64, respectively. In the training stage, the activation function is Relu, the learning rate is 0.0001, and a dropout with a coefficient of 0.5 is used to prevent the model from overfitting. Applying the proposed acoustic representation with TCN yields a higher PIA ranging from 3.4% to 21.2%. Overall, it is also useful to briefly review classification performance figures reported on the IEMOCAP corpus by other works. Although the numbers cannot be directly compared due to factors such as different data partitioning, they are still useful for general benchmarking. Unless otherwise specified, the results cited here are achieved for four basic emotional categories with different acoustic representation schemes. Again, for the IEMOCAP corpus, the classification accuracy by the proposed approach is fairly higher than that obtained by previous works in [146–152], giving the highest PIA reaching up to 17.1%.

The well-established descriptive framework that uses categorical emotions offers intuitive emotion descriptions and is widely used [49]. The combination of basic emotional labels can also serve as a convenient representation of the universal emotional space [1]. There is an increasing trend in studying emotional states by dimensional approaches [19, 26, 43, 66, 157, 158]. A dimensional framework enables for gradual change within the same emotion as well as a transition between emotional states. In this study, we predict the three continuous emotional primitives of valence, arousal, and dominance in the IEMOCAP corpus. As shown in Table IV, the performance tendency by means of MTAR, VEE, and AEP is the same as the trend observed in Table III. The MTAR appears to be more suitable for SER relative to the two individual feature subsets despite computational models, indicating the effective representation of combined music theory-inspired VEE and AEP indeed contributes to recognition performance. It also is noticed from Table IV that the proposed MTAR is always superior to the spectrogram, Mel-spectrogram, MFCCs, and VGGish representations on regression performance, regardless of the computational models. It improves the CCC values for all three-dimensional emotion primitives of valence, arousal, and dominance, reaching the highest CCCs of up to 0.474,

TABLE IV
REGRESSION RESULTS OF DIMENSIONAL EMOTIONS BY DIFFERENT ACOUSTIC REPRESENTATIONS AND THE STATE-OF-THE-ART WORKS.

Methods	Features	Valence	Arousal	Dominance	UA	PIA
BiGRU	**MTAR**	**0.464**	0.684	0.505	**0.551**	-
	VEE	0.326	**0.688**	**0.527**	0.513	7.8%
	AEP	0.436	0.641	0.481	0.519	6.7%
	Spectrogram	0.403	0.580	0.420	0.468	15.7%
	MelSpectrogram	0.435	0.626	0.460	0.507	8.9%
	MFCCs	0.460	0.607	0.463	0.510	8.4%
	VGGish	0.373	0.667	0.500	0.513	7.8%
CNN	**MTAR**	**0.446**	**0.674**	**0.519**	**0.546**	-
	VEE	0.252	0.643	0.468	0.454	16.8%
	AEP	0.438	0.650	0.489	0.525	4.4%
	Spectrogram	0.390	0.644	0.482	0.505	8.3%
	MelSpectrogram	0.440	0.659	0.493	0.531	3.3%
	MFCCs	0.437	0.657	0.478	0.524	4.7%
	VGGish	0.345	0.664	0.470	0.493	10.5%
TCN	MTAR	**0.474**	**0.690**	**0.521**	**0.562**	-
	VEE	0.290	0.626	0.481	0.465	18.1%
	AEP	0.455	0.665	0.497	0.539	5.0%
	Spectrogram	0.423	0.651	0.488	0.521	8.6%
	MelSpectrogram	0.462	0.666	0.497	0.542	4.4%
	MFCCs	0.465	0.663	0.500	0.543	4.2%
	VGGish	0.351	0.674	0.485	0.503	11.7%
SVM	InterSp09	0.381	0.650	0.483	0.505	11.5%
	InterSp10	0.433	0.671	0.506	0.537	5.4%
	InterSp11	0.391	0.672	0.505	0.523	8.2%
	InterSp12	0.387	0.674	0.509	0.523	8.2%
	InterSp13	0.390	0.674	0.504	0.523	8.2%
	GeMAPs	0.317	0.641	0.466	0.475	16.6%
DNN	InterSp09	0.351	0.557	0.422	0.443	21.3%
	InterSp10	0.384	0.549	0.392	0.442	21.5%
	InterSp11	0.397	0.558	0.399	0.451	20.1%
	InterSp12	0.417	0.589	0.426	0.477	16.1%
	InterSp13	0.405	0.585	0.414	0.468	17.6%
	GeMAPs	0.373	0.565	0.422	0.453	19.8%
MLP [157]	GeMAPs+HSFs	0.316	0.599	0.473	0.463	18.4%
LSTM [158]	GeMAPs	0.192	0.553	0.456	0.400	26.9%
	pAA	0.183	0.577	0.444	0.401	26.8%
LSTM [159]	GeMAPs	0.168	0.486	0.442	0.365	30.9%
	HSF1	0.206	0.526	0.442	0.391	28.0%
	HFS2	0.204	0.544	0.442	0.397	27.3%

0.690, and 0.521, respectively. The PIA values are clearly notable regardless of the BiGRU, CNN, and TCN models applied. In addition, the commonly-used brute-forced parameter sets in INTERSPEECH from 2009 to 2016 were compared by means of the SVC and DNN models, with the same configuration for the categorical SER task, but the radial basis function (RBF) kernel is used in the SVC configuration, and the cross-entropy loss function is replaced by the mean square error in the DNN scenario. Again, the advantage of the proposed approach over these six INTERSPEECH feature sets is in predicting all these emotion dimensions. For further analysis, the other studies targeting dimensional SER have also been included in Table IV. By comparison, a positive result of ours is that the overall CCC reached up to 0.561 in a TCN-based SER scenario, resulting in a PIA of [18.4%, 30.9%] over the previous attempts [157–159]. Notably, it is interesting that we achieved better performance on valence regression because it is a challenging task in most dimensional SER studies [19, 65, 66], whereas this valence dimension is significantly important to distinct emotional pairs with the same arousal level, such as anger vs. happiness classification. It is promising for emotional AI of call-centers applications.

V. Conclusion

This work presents a novel acoustic representation for the recognition of human emotions in speech. To this end, we introduced two new attempts to design acoustic representation of emotional speech regarding vocal emotion expression and auditory emotion perception processes, respectively, via investigating music theory contents.

The proposed acoustic representation of emotion was evaluated on the IEMOCAP corpus to i) classify four emotional categories and ii) predict three emotional dimensions. The individual acoustic representation of spectrogram, Mel-spectrogram, MFCCs, and VGGish, as well as six predefined feature sets widely used in the INTERSPEECH Challenges on Emotion and Paralinguistics from 2009 to 2016, were included for reference. Experiments were done under conditions of LOSO-CV. The proposed acoustic representation consistently provided a better performance under conditions of three different computational models dominated by BiGRU, CNN, and TCN, yielding an average classification accuracy of 61.9%, 62.5%, and 63.2%, and an average regression CCCs over valence, arousal, and dominance by 0.551, 0.546, and 0.561, respectively. These results showed the benefit of using the MTAR. Moreover, we reviewed the literature as general benchmarks. Comparing with the related work targeting recognition of the same emotional corpus, the proposed acoustic representation fairly showed improved results for both categorical and dimensional SER tasks. In particular, the new MTAR appears to offer potential for the prediction of the valence dimension, which is a significantly challenging issue in literature and even poorly evaluated by human listeners.

Numerous studies over the past few decades have tried recognizing emotions in speech. Increasing efforts have been made to design outstanding features that best reflect and differentiate specific emotional information. A series of prior studies have reported the overlapping neurophysiological, cognitive, and perceptual processes in common between music and speech. Concerning the processing of human vocal emotion expression and auditory emotion perception, this study achieved a new computational acoustic representation by studying musical contents in speech. With possible refinement in future work, the performance of MTAR could be further improved. Hence, further research on the use of music theory-inspired acoustics for SER can be beneficial.

Acknowledgment

This research was supported by the Key research and development program of Hainan province (ZDYF2021GXJS017), the National Natural Science Foundation of China (82160345 and 62201571), the Key science and technology plan project of Haikou (2011-016).

References

[1] Paul Ed Ekman and Richard J Davidson, *The nature of emotion: Fundamental questions.*, Oxford University Press, 1994.
[2] Jaak Panksepp, *Affective neuroscience: The foundations of human and animal emotions*, Oxford university press, 2004.
[3] Daniel L Schacter, "Psychology second edition, 41 madison avenue," *New York, NY*, vol. 10010, pp. 310, 2011.
[4] Peggy A Thoits, "The sociology of emotions," *Annual review of sociology*, vol. 15, no. 1, pp. 317–342, 1989.
[5] Timothy D Wilson and Elizabeth W Dunn, "Self-knowledge: Its limits, value, and potential for improvement," *Annu. Rev. Psychol.*, vol. 55, pp. 493–518, 2004.
[6] Lisa Feldman Barrett and James A Russell, *The psychological construction of emotion*, Guilford Publications, 2014.
[7] Norbert Schwarz, *Feelings as information: Informational and motivational functions of affective states.*, The Guilford Press, 1990.
[8] Teresa E Seeman, Tina M Lusignolo, Marilyn Albert, and Lisa Berkman, "Social relationships, social support, and patterns of cognitive aging in healthy, high-functioning older adults: Macarthur studies of successful aging.," *Health psychology*, vol. 20, no. 4, pp. 243, 2001.
[9] D Caroline Blanchard, April L Hynd, Karl A Minke, Tiffanie Minemoto, and Robert J Blanchard, "Human defensive behaviors to threat scenarios show parallels to fear-and anxiety-related defense patterns of non-human mammals," *Neuroscience & Biobehavioral Reviews*, vol. 25, no. 7-8, pp. 761–770, 2001.
[10] Gerald L Kooyman, *Diverse divers: physiology and behavior*, vol. 23, Springer Science & Business Media, 2012.
[11] Amber Haque, "Psychology from islamic perspective: Contributions of early muslim scholars and challenges to contemporary muslim psychologists," *Journal of religion and health*, vol. 43, no. 4, pp. 357–377, 2004.
[12] Graeme J Taylor, "Alexithymia: concept, measurement, and implications for treatment.," *The American journal of psychiatry*, 1984.
[13] George Mandler, *Mind and body: Psychology of emotion and stress*, WW Norton & Company Incorporated, 1984.
[14] Paul Edmund Griffiths and Andrea Scarantino, "Emotions in the wild: The situated perspective on emotion," 2005.
[15] Yana Suchy, *Clinical neuropsychology of emotion*, Guilford Press, 2011.
[16] Jerome Kagan et al., *What is emotion?: History, measures, and meanings*, Yale University Press, 2007.
[17] Kenneth Tobin, Stephen M Ritchie, Jennifer L Oakley, Victoria Mergard, and Peter Hudson, "Relationships between emotional climate and the fluency of classroom interactions," *Learning Environments Research*, vol. 16, no. 1, pp. 71–89, 2013.
[18] Rosalind W Picard, *Affective computing*, MIT press, 2000.
[19] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al., "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer speech & language*, vol. 29, no. 1, pp. 100–131, 2015.
[20] William James, "What is emotion? 1884.," 1948.
[21] Jianhua Tao and T Tieniu, "Affective computing: A review. affective computing and intelligent interaction. lncs 3784," *Springer*, vol. 981, pp. 995, 2005.
[22] Ognjen Rudovic, Nicolas Tobis, Sebastian Kaltwang, Björn Schuller, Daniel Rueckert, Jeffrey F Cohn, and Rosalind W Picard, "Personalized

federated deep learning for pain estimation from face images," *arXiv preprint arXiv:2101.04800*, 2021.
[23] Jinni Harrigan, Robert Rosenthal, Klaus R Scherer, and Klaus Scherer, *New handbook of methods in nonverbal behavior research*, Oxford University Press, 2008.
[24] Angeliki Metallinou, Athanassios Katsamanis, Yun Wang, and Shrikanth Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *ICASSP*. IEEE, 2011, pp. 2288–2291.
[25] Ginevra Castellano, Loic Kessous, and George Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*, pp. 92–103. Springer, 2008.
[26] Björn 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.
[27] Pritam Khan, Priyesh Ranjan, and Sudhir Kumar, "At2gru: A human emotion recognition model with mitigated device heterogeneity," *IEEE Transactions on Affective Computing*, 2021.
[28] Magdalena Sandner, Peter Zeier, Giannis Lois, and Michèle Wessa, "Cognitive emotion regulation withstands the stress test: An fmri study on the effect of acute stress on distraction and reappraisal," *Neuropsychologia*, vol. 157, pp. 107876, 2021.
[29] Sahar Karimi Shahraki and Mahdi Khezri, "Identification of attention deficit hyperactivity disorder patients using wavelet-based features of eeg signals," *Journal of Intelligent Procedures in Electrical Technology*, vol. 12, no. 47, pp. 1–11, 2021.
[30] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.
[31] Richard Yonck, *Heart of the machine: Our future in a world of artificial emotional intelligence*, Arcade, 2020.
[32] Kiel Gilleade, Alan Dix, and Jen Allanson, "Affective videogames and modes of affective gaming: assist me, challenge me, emote me," *DiGRA 2005: Changing Views–Worlds in Play.*, 2005.
[33] Joris H Janssen, Egon L Van Den Broek, and Joyce HDM Westerink, "Tune in to your emotions: a robust personalized affective music player," *User Modeling and User-Adapted Interaction*, vol. 22, no. 3, pp. 255–279, 2012.
[34] Abdelrahman El-Amin, Ahmed Attia, Omar Hammad, Osama Nasr, Osama Ghozlan, Remon Raouf, Ahmed M Hamed, Hany Eldawlatly, Magdy El-Moursy, and Seif Eldawlatly, "Brain-in-car: A brain activity-based emotion recognition embedded system for automotive," in *2019 ICVES*. IEEE, 2019, pp. 1–5.
[35] Petri Laukka and Hillary Anger Elfenbein, "Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis," *Emotion Review*, vol. 13, no. 1, pp. 3–11, 2021.
[36] Johan Sundberg, Gláucia Laís Salomão, and Klaus R Scherer, "Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation," *Journal of Voice*, vol. 35, no. 1, pp. 52–60, 2021.
[37] Mehmet Berkehan Akçay and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.
[38] Sung-Woo Byun and Seok-Pil Lee, "A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms," *Applied Sciences*, vol. 11, no. 4, pp. 1890, 2021.
[39] Jiahong Yuan, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church, "The role of phonetic units in speech emotion recognition," *arXiv preprint arXiv:2108.01132*, 2021.
[40] Patrik N Juslin and Petri Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological bulletin*, vol. 129, no. 5, pp. 770, 2003.
[41] Klaus R Scherer, "Voices of power, passion, and personality.," in *INTERSPEECH*, 2015.
[42] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang, "Speechin: A smart necklace for silent speech recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–23, 2021.
[43] Yongwei Li, Jianhua Tao, Donna Erickson, Bin Liu, and Masato Akagi, "f_0-noise-robust glottal source and vocal tract analysis based on arx-lf model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3375–3383, 2021.
[44] Johan Sundberg, Sona Patel, Eva Bjorkner, and Klaus R Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011.
[45] Kurt Hammerschmidt and Uwe Jürgens, "Acoustical correlates of affective prosody," *Journal of voice*, vol. 21, no. 5, pp. 531–540, 2007.
[46] Florian Eyben, Felix Weninger, and Björn Schuller, "Affect recognition in real-life acoustic conditions-a new perspective on feature selection," in *2013 INTERSPEECH 2013*, 2013.
[47] Penny Bergman, Daniel Västfjäll, Ana Tajadura-Jiménez, and Erkin Asutay, "Auditory-induced emotion mediates perceptual categorization of everyday sounds," *Frontiers in psychology*, vol. 7, pp. 1565, 2016.
[48] Lin Jiang, Ping Tan, Junfeng Yang, Xingbao Liu, and Chao Wang, "Speech emotion recognition using emotion perception spectral feature," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 11, pp. e5427, 2021.
[49] Siqing Wu, Tiago H Falk, and Wai-Yip Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.
[50] Masashi Unoki and Zhi Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.
[51] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge," 2009.
[52] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.
[53] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski, "The interspeech 2011 speaker state challenge," 2011.
[54] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al., "The interspeech 2012 speaker trait challenge," in *INTERSPEECH 2012, Portland, OR, USA*, 2012.
[55] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang, "The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking," in *2014 INTERSPEECH, Singapore*, 2014.
[56] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
[57] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
[58] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al., "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *2016 Interspeech, Vols 1-5*, 2016, pp. 2001–2005.
[59] Aaron Keesing, Yun Sing Koh, and Michael Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," *Proc. Interspeech 2021*, pp. 3415–3419, 2021.
[60] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.
[61] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *2017 ICASSP*. IEEE, 2017, pp. 131–135.
[62] Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan, "Ordinal learning for emotion recognition in customer service calls," in *2020 ICASSP*. IEEE, 2020, pp. 6494–6498.
[63] Kyoung Ju Noh, Chi Yoon Jeong, Jiyoun Lim, Seungeun Chung, Gague Kim, Jeong Mook Lim, and Hyuntae Jeong, "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets," *Sensors*, vol. 21, no. 5, pp. 1579, 2021.
[64] T Mani Kumar, Enrique Sanchez, Georgios Tzimiropoulos, Timo Giesbrecht, and Michel Valstar, "Stochastic process regression for cross-cultural speech emotion recognition," *Proc. Interspeech 2021*, pp. 3390–3394, 2021.
[65] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[66] Xingfeng Li and Masato Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Communication*, vol. 110, pp. 1–12, 2019.

[67] Charles Darwin, *The expression of the emotions in man and animals*, Kartindo. com, 1948.

[68] Leonard Bernstein, *The unanswered question: Six talks at Harvard*, vol. 33, Harvard University Press, 1976.

[69] Klaus R Scherer, "Vocal affect expression: a review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, pp. 143, 1986.

[70] Alf Gabrielsson and Patrik N Juslin, *Emotional expression in music.*, Oxford University Press, 2003.

[71] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.

[72] Aniruddh D Patel, *Music, language, and the brain*, Oxford university press, 2010.

[73] Lutz Jäncke, "The relationship between music and language," *Frontiers in psychology*, vol. 3, pp. 123, 2012.

[74] Dianna Vidas, Genevieve A Dingle, and Nicole L Nelson, "Children's recognition of emotion in music and speech," *Music & Science*, vol. 1, pp. 2059204318762650, 2018.

[75] Peter Kivy, *Sound sentiment: An essay on the musical emotions, including the complete text of the corded shell*, Temple University Press, 1989.

[76] Ray Jackendoff and Fred Lerdahl, "A grammatical parallel between music and language," in *Music, mind, and brain*, pp. 83–117. Springer, 1982.

[77] Klaus R Scherer, Johan Sundberg, Bernardino Fantini, Stéphanie Trznadel, and Florian Eyben, "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1805–1815, 2017.

[78] Klaus R Scherer, "Expression of emotion in voice and music," *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[79] Sebastian Jentschke, *The relationship between music and language*, Oxford University Press, Oxford, UK,, 2015.

[80] Sandra E Trehub and Takayuki Nakata, "Emotion and music in infancy," *Musicae scientiae*, vol. 5, no. 1_suppl, pp. 37–61, 2001.

[81] Disa A Sauter, Charlotte Panattoni, and Francesca Happé, "Children's recognition of emotions from vocal cues," *British Journal of Developmental Psychology*, vol. 31, no. 1, pp. 97–113, 2013.

[82] Adam J Lonsdale and Adrian C North, "Why do we listen to music? a uses and gratifications analysis," *British journal of psychology*, vol. 102, no. 1, pp. 108–134, 2011.

[83] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[84] Deborah Ross, Jonathan Choi, and Dale Purves, "Musical intervals in speech," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9852–9857, 2007.

[85] Josh H McDermott, Michael V Keebler, Christophe Micheyl, and Andrew J Oxenham, "Musical intervals and relative pitch: Frequency resolution, not interval resolution, is special," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1943–1951, 2010.

[86] Didier Grandjean, "Brain networks of emotional prosody processing," *Emotion Review*, vol. 13, no. 1, pp. 34–43, 2021.

[87] Wayne Slawson, *Sound color*, Yank Gulch Music, 1985.

[88] Fred Lerdahl, "Timbral hierarchies," *Contemporary Music Review*, vol. 2, no. 1, pp. 135–160, 1987.

[89] Patrik N Juslin and John Sloboda, *Handbook of music and emotion: Theory, research, applications*, Oxford University Press, 2011.

[90] Simone Falk, Tamara Rathcke, and Simone Dalla Bella, "When speech sounds like music.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 4, pp. 1491, 2014.

[91] Siu-Lan Tan, Peter Pfordresher, and Rom Harré, *Psychology of Music: From Sound to Significance Second Edition*, Routledge, 2017.

[92] Keith Swanwick, *Music, mind and education*, Routledge, 2003.

[93] Cengiz Kaygusuz and Julian Zuluaga, "Impact of intervals on the emotional effect in western music," *arXiv preprint arXiv:1812.04723*, 2018.

[94] William Forde Thompson, E Glenn Schellenberg, and Gabriela Husain, "Decoding speech prosody: Do music lessons help?," *Emotion*, vol. 4, no. 1, pp. 46, 2004.

[95] Ken'ichi Miyazaki, "Absolute pitch identification: Effects of timbre and pitch region," *Music perception*, vol. 7, no. 1, pp. 1–14, 1989.

[96] Mayumi Adachi and Sandra E Trehub, "Children's expression of emotion in song," *Psychology of music*, vol. 26, no. 2, pp. 133–153, 1998.

[97] Lauren Stewart, Tobias Overath, Jason D Warren, Jessica M Foxton, and Timothy D Griffiths, "fmri evidence for a cortical hierarchy of pitch pattern processing," *PLoS One*, vol. 3, no. 1, pp. e1470, 2008.

[98] Takashi Fujisawa, Kazuaki Takami, and Norman D Cook, "On the role of pitch intervals in the perception of emotional speech," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[99] Bin Yang and Marko Lugger, "Emotion recognition from speech signals using new harmony features," *Signal processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[100] Fred Lerdahl and Ray S Jackendoff, *A Generative Theory of Tonal Music, reissue, with a new preface*, MIT press, 1996.

[101] Josh H McDermott and Andrew J Oxenham, "Music perception, pitch, and the auditory system," *Current opinion in neurobiology*, vol. 18, no. 4, pp. 452–463, 2008.

[102] AJ Hudspeth, Y Choe, AD Mehta, and P Martin, "Putting ion channels to work: mechanoelectrical transduction, adaptation, and amplification by hair cells," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11765–11772, 2000.

[103] Nils Lennart Wallin, Björn Merker, and Steven Brown, *The origins of music*, MIT press, 2000.

[104] Steven Brown, "Are music and language homologues?," *Annals of the New York Academy of Sciences*, vol. 930, no. 1, pp. 372–374, 2001.

[105] Aleksey Nikolsky, "Commentary: the 'musilanguage'model of language evolution," *Frontiers in psychology*, vol. 9, pp. 75, 2018.

[106] Manfred Clynes, *Music, mind, and brain: The neuropsychology of music*, Springer Science & Business Media, 2013.

[107] Ruth Lesser, "Verbal comprehension in aphasia: An english version of three italian tests," *Cortex*, vol. 10, no. 3, pp. 247–263, 1974.

[108] Carol L Krumhansl, "The cognition of tonality–as we know it today," *Journal of New Music Research*, vol. 33, no. 3, pp. 253–268, 2004.

[109] Barbara Tillmann, "Implicit investigations of tonal knowledge in nonmusician listeners.," in *The Neurosciences and Music II: From Perception to Performance, May, 2005, Leipzig, Germany; The work in this volume is the result of the aforementioned conference.* New York Academy of Sciences, 2005.

[110] Peter B Denes, Peter Denes, and Elliot Pinson, *The speech chain*, Macmillan, 1993.

[111] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Machine speech chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020.

[112] SeyyedPooya HekmatiAthar and Mohd Anwar, "Music embedding: A tool for incorporating music theory into computational music applications," *arXiv preprint arXiv:2104.11880*, 2021.

[113] Sandrine Vieillard, Isabelle Peretz, Nathalie Gosselin, Stéphanie Khalfa, Lise Gagnon, and Bernard Bouchard, "Happy, sad, scary and peaceful musical excerpts for research on emotions," *Cognition & Emotion*, vol. 22, no. 4, pp. 720–752, 2008.

[114] Eunjin Choi, Yoonjin Chung, Seolhee Lee, JongIk Jeon, Taegyun Kwon, and Juhan Nam, "Ym2413-mdb: A multi-instrumental fm video game music dataset with emotion annotations," *arXiv preprint arXiv:2211.07131*, 2022.

[115] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, 2020.

[116] António Pedro Oliveira and Amílcar Cardoso, "A musical system for emotional expression," *Knowledge-Based Systems*, vol. 23, no. 8, pp. 901–913, 2010.

[117] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *arXiv preprint arXiv:2001.00459*, 2019.

[118] Stephanie M Stalinski and E Glenn Schellenberg, "Music cognition: a developmental perspective," *Topics in Cognitive Science*, vol. 4, no. 4, pp. 485–497, 2012.

[119] MIDI Complete, "1.0 detailed specification," *MIDI Manufacturers Association Inc*, 1996.

[120] Matthias Thiemel, "Dynamics. grove music online," 2001.

[121] Anders Elowsson and Anders Friberg, "Predicting the perception of performed dynamics in music audio with ensemble learning," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2224–2242, 2017.

[122] James F Kaiser, "On a simple algorithm to calculate the'energy'of a signal," in *International conference on acoustics, speech, and signal processing*. IEEE, 1990, pp. 381–384.

[123] James F Kaiser, "Some useful properties of teager's energy operators,"

in *1993 IEEE international conference on acoustics, speech, and signal processing*. IEEE, 1993, vol. 3, pp. 149–152.
[124] Ebenezer Prout, *Harmony: its theory and practice*, Cambridge University Press, 2011.
[125] Stanley Sadie and John Tyrrell, *Dictionary of music and musicians*, New York: Oxford University Press. Yónatan Sánchez, 2001.
[126] Imre Lahdelma and Tuomas Eerola, "Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.
[127] Emili Renard i Vallet et al., "sonancia: una clarificación conceptual," 2016.
[128] Haye Hinrichsen, "Revising the musical equal temperament," *Revista Brasileira de Ensino de Física*, vol. 38, no. 1, 2016.
[129] Willi Apel, *The Harvard dictionary of music*, Harvard University Press, 2003.
[130] Glenn Spring and Jere Hutcheson, *Musical form and analysis: Time, pattern, proportion*, Waveland Press, 2013.
[131] Alexander John Ellis, *On the musical scales of various nations*, Journal of the Society of arts, 1885.
[132] Catherine M Warrier and Robert J Zatorre, "Influence of tonal context and timbral variation on perception of pitch," *Perception & psychophysics*, vol. 64, no. 2, pp. 198–207, 2002.
[133] Dave Benson, *Music: A mathematical offering*, Cambridge University Press, 2006.
[134] Michel Marie Deza and Elena Deza, "Image and audio distances," in *Encyclopedia of Distances*, pp. 387–411. Springer, 2014.
[135] David Ryan, "Mathematical harmony analysis," *arXiv preprint arXiv:1603.08904*, 2016.
[136] Eberhard Zwicker and Hugo Fastl, *Psychoacoustics: Facts and models*, vol. 22, Springer Science & Business Media, 2013.
[137] Meinard Muller and Sebastian Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 649–662, 2010.
[138] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *2013 INTERSPEECH*, 2013.
[139] Alex Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132306, 2020.
[140] Dimitrios Kollias and Stefanos Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 595–606, 2020.
[141] Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, and Jiahui Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, pp. 11–19, 2020.
[142] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever, "An empirical exploration of recurrent network architectures," in *International conference on machine learning*. PMLR, 2015, pp. 2342–2350.
[143] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[144] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 ICASSP*. IEEE, 2016, pp. 5200–5204.
[145] Wootaek Lim, Daeyoung Jang, and Taejin Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 APSIPA*. IEEE, 2016, pp. 1–4.
[146] Michael Neumann and Ngoc Thang Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *2019 ICASSP*. IEEE, 2019, pp. 7390–7394.
[147] DN Krishna and Ankita Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks.," in *Interspeech*, 2020, pp. 4243–4247.
[148] Lili Guo, Longbiao Wang, Chenglin Xu, Jianwu Dang, Eng Siong Chng, and Haizhou Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *2021 ICASSP*. IEEE, 2021, pp. 6304–6308.
[149] Andreas Triantafyllopoulos, Shuo Liu, and Björn W Schuller, "Deep speaker conditioning for speech emotion recognition," in *2021 ICME*. IEEE, 2021, pp. 1–6.
[150] Ruichen Li, Jinming Zhao, and Qin Jin, "Speech emotion recognition via multi-level cross-modal distillation," *Proc. Interspeech 2021*, pp. 4488–4492, 2021.
[151] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *2020 ICASSP*. IEEE, 2020, pp. 7144–7148.
[152] Xixin Wu, Shoukang Hu, Zhiyong Wu, Xunying Liu, and Helen Meng, "Neural architecture search for speech emotion recognition," in *2022 ICASSP*. IEEE, 2022, pp. 6902–6906.
[153] Riccardo Williams, "Anger as a basic emotion and its role in personality building and pathological growth: The neuroscientific, developmental and clinical perspectives," *Frontiers in psychology*, vol. 8, pp. 1950, 2017.
[154] Patrik N Juslin and Petri Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of new music research*, vol. 33, no. 3, pp. 217–238, 2004.
[155] Casper J Albers and Laura F Bringmann, "Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model," *European Journal of Psychological Assessment*, 2020.
[156] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[157] Bagus Tris Atmaja and Masato Akagi, "Deep multilayer perceptrons for dimensional speech emotion recognition," in *2020 APSIPA*. IEEE, 2020, pp. 325–331.
[158] Bagus Tris Atmaja and Masato Akagi, "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm," *Speech Communication*, vol. 126, pp. 9–21, 2021.
[159] Bagus Tris Atmaja and Masato Akagi, "Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition," in *Journal of Physics: Conference Series*. IOP Publishing, 2021, vol. 1896, p. 012004.

VI. Biography Section

Xingfeng Li received the M.S. degrees in software engineering and information science from Tianjin University, China, and Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2016, respectively, and the Ph.D. degree in information science from JAIST, in 2019. Since 2022, he has been on the faculty of the School of Computer Science and Technology, Hainan University, Haikou, China, and is currently an Associate Professor. His research interests are affective computing, speech processing, and speech perception, emphasizing how para/non-linguistic information (speech emotion) impacts spoken communication. He was a member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). He was awarded the Chinese Government Scholarship to sponsor his doctoral study in JAIST from 2016 to 2019 and the Best Oral Paper from the Oriental COCOSDA in 2015.

Xiaohan Shi received the B.E. degree in computer science and technology from Shaanxi University of Technology, China, in 2017, and the M.S. degree in information science from Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2020. He is currently pursuing a Ph.D. degree at Nagoya University, Nagoya, Japan. He is now an Interdisciplinary Frontier Next Generation Researcher with Japan Science and Technology Agency (JST) from 2022 to 2025, a member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). His research interests are affective computing and speech processing.

Desheng Hu received the B.E. degree in communication engineering from the University of Jinan, China, in 2018, and the M.S. degree in 2021 in electronics and communication engineering from Taiyuan University of Technology, Shanxi, China. He is now an engineer at Hithink RoyalFlush. His research interests include affective computing, speech synthesis, and deep learning.

Yongwei Li received the M.S. and Ph.D. degrees in information science form the Japan Advanced Institute of Science and Technology, Nomi, Japan, in 2014 and 2018, respectively. He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,Bejing, China. His current research interests include modeling of speech production, voice quality, and speech emotion recognition and synthesis. He is an associate editor of Acoustical Science and Technology.

Qingchen Zhang (Senior Member, IEEE) is currently a professor at Hainan University, China. He obtained his Ph.D degree from Dalian University of Technology, China, in 2015. His current research is smart medicine. He has published around 50 referred papers on top IEEE/ACM Transactions/Journals/Magazines. He was awarded for IEEE TCCPS Most Influential Paper Award (2020), IEEE SCSTC Most Influential Paper Award (2019), and IEEE TCSC Award for Excellence in Scalable Computing (Early Career, 2018). Dr. Zhang is serving as an associate editor of Journal of Ambient Intelligence and Humanized Computing and Journal of Circuits, Systems and Computers. He served as a program chair of some international conferences such as IEEE 2020 Cybermatics Congress and IEEE 2018 International Conference on Internet of Things.

Zhengxia Wang received the B.E. Degree from Chongqing Jiaotong University, China, in 2001, the M.S. and PhD degrees in computer science and technology from Chongqing University, China, in 2004 and 2009, respectively. She is currently a professor with Hainan University China. Her research interests include machine learning, data mining and image processing.

Masashi Unoki (Member, IEEE) received his M.S. and Ph.D. in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His main research interests are inauditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher from 1999-2000, and he was a visiting research associate at the Centre for the Neural Basis of Hearing (CNBH) in the Department of Physiology at the University of Cambridge from 2000 to 2001. He has been on the faculty of the School of Information Science at JAIST since 2001 and a full professor. Now, he is a Councilor of JAIST. Dr. Unoki received the Sato Prize from the Acoustical Society of Japan (ASJ) in 1999, 2010, and 2013 for Outstanding Papers and Best Paper Award from the Institute of Electronics, Information and Communication Engineers in 2017. Currently, he is an associate editor of Applied Acoustics.

Masato Akagi (Life Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been a Faculty Member of the School of Information Science, JAIST, and is currently a Professor Emeritus. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech. Dr. Akagi was the recipient of the IEICE Excellent Paper Award from the IEICE in 1987, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the Acoustical Society of Japan in 1998, 2005, 2010 and 2011.