BEES: A New Acoustic Task for Blended Emotion Estimation in Speech

Xingfeng Li[1], Xiaohan Shi[2], Yuke Si[3], Zilong Zhang[4], Feifei Cui[5],
Yongwei Li[6], Yang Liu[7], Masashi Unoki[8], Masato Akagi[9]
[1] City University of Macau, Macau. E-mail: xfli@cityu.edu.mo
[2] Nagoya University, Nagoya. E-mail: xiaohan.shi@g.sp.m.is.nagoya-u.ac.jp
[3] China Mobile Research Institute, Beijing. E-mail: siyuke@163.com
[4, 5] Hainan University, Haikou. E-mail: {zhangzilong, feifeicui} @hainanu.edu.cn
[6] Institute of Psychology, Chinese Academy of Sciences, Beijing. E-mail: liyw@psych.ac.cn
[7] Qingdao University of Science and Technology, Qingdao. E-mail: yangliu@qust.edu.cn
[8, 9] Japan Advanced Institute of Science and Technology, Nomi. E-mail: {unoki, akagi} @jaist.ac.jp

***Abstract*—Acoustic emotion recognition (AER) is a challenging yet crucial area in affective computing. Despite the clear importance of perceiving vocal emotion blends with varying intensities for understanding the emotional complexity of everyday life, most existing studies remain limited to single-label emotion classification tasks. This paper introduces a novel acoustic task for blended emotion estimation in speech (BEES), which aims to capture the intricate emotional distributions of co-existing emotions with varied intensities in voices. The BEES task stands out by defining a precise reference for the emotion intensity of vocal blends, introducing innovative evaluation metrics and providing robust baseline results. Additionally, our study highlights the significant impact of self-supervised learning representations, voice naturalness, and gender on BEES performance. By addressing the limitations of traditional AER methods and offering a comprehensive framework for analyzing blended emotions, this work paves the way for advanced computational AER systems capable of handling the complexity of vocal emotion blends.**

I. Introduction

Emotion is a psychological state brought on by neurophysiological changes, variously associated with thoughts, feelings, and behavioral responses [1]. It can function to communicate positively important information to individuals in significant external events, such as values and ethics [2]. Also, it is sometimes internally regarded as part of a mental illness and thus possibly of negative value, for instance, anxiety or depression [3]. There is widespread evidence supporting that emotion is of vital importance for social competence, decision-making, and well-being [4]. Within its research area of interest, acoustic emotion recognition (AER) aims to give machines emotional intelligence underlying how humans feel emotionally towards their voice has drawn great attention and could be beneficial to wide applications such as customer care service, human-machine interaction, and many others [5], [6].

Most of the existing works toward AER can be regarded as a single-label classification task, assuming that each voice stimulus only evokes a dominant emotion [7]–[9]. Despite substantial progress made in this area, these works however over-simplify the complexity of human emotions and neglect the ambiguity and subjectivity that lie in them. Historically and today, advances in psychology research substantiate the existence of the expression and perception of emotion-blends, indicating that humans can frequently experience more than one emotion at the same time [10]–[13]. For instance, Juslin et al. (2021) conducted a study capturing naturalistic vocal expressions in a field setting [14]. Speakers themselves provided the ground truth of emotional expressions in each recording, revealing that 41% of the recordings depicted situations where blended emotions were experienced. Also, other nonverbal communication research has further investigated various types of blended emotional experiences such as the co-activation of happiness and sadness [13], fear and happiness [15], disgust and amusement [16], as well as hope and fear [17]. It is grounded that treating AER as a single-label classification problem is oversimplified and lacks practical applicability.

To address acoustical emotion-blends estimation, multi-label learning has been extensively investigated, which usually selects a threshold for the output of a classifier, then labels emotions with scores higher than the threshold as existing and the others as not existing [18], [19]. Unfortunately, these methods failed to estimate the intensity of each specific emotion. On the other hand, recent advancements in machine learning suggest the existence of label distribution learning which can be promising to represent the degree to which each label describes the instance [20], [21]. For any instance, the sum of the description degrees of all labels is one, indicating a full description of this instance. Inspired by the above studies, we first introduce an acoustic task for BEES, assuming each voice stimulus contains a blend of multiple emotions with varying intensities. Specifically, we label each stimulus by an emotion vector, where each element corresponds to a specific emotion and the value of each element is the intensity of that emotion. In this context, the emotion vectors can be interpreted as emotion distributions, and BEES aims to map human voices to their corresponding emotion distributions by maximizing the similarities between distributions obtained by human evaluations and system estimations.

To our knowledge, this paper takes one step beyond current AER algorithms and is the first attempt at BEES to investigate machine learning-based emotion-blends estimation in human voices. The main contributions of this paper are multi-fold: (1) we introduce an important human assessment criterion to produce blended emotion references in voice stimulus, (2) we identify efficient evaluation metrics to assess the BEES performance, (3) we discuss the effects of acoustic features, emotion production styles, and genders on the BEES performance. The remainder of this article is structured as follows: Section 2 outlines the utilized data and human assessment methodology for identifying blended emotion references in human voices. Section 3 presents the foundational acoustic features of self-supervised learning representations (SSLs), estimation methodologies, and evaluation metrics applied to the tasks of BEES. Section 4 provides an elaborate breakdown of the BEES results along with corresponding discussions. Finally, Section 5 presents pivotal conclusions drawn from this study.

II. Blended Emotion Estimation in Speech

A. Emotional Speech Database

We use the popular IEMOCAP corpus for this study for the following four reasons [22]. First, the evaluators annotating the IEMOCAP corpus were allowed to tag more than one emotional category per speech stimulus, to account for mixtures of emotions such as frustration and anger, etc, which are frequently observed in real-life scenarios. Secondly, IEMOCAP defines N-category labels by carefully balancing the trade-off between the number of emotion categories, providing a more accurate and detailed emotional description with higher agreement among evaluators. It contains nine categorical labels of emotion, including anger (ang.), sadness (sad.), happiness (hap.), disgust (dis.), fear (fea.), and surprise (sur.) which are known as basic emotions, plus frustration (fru.), excited (exc.), and neutral (neu.) states. Third, this corpus was deliberately chosen to contain two different emotion production styles. The spontaneous speech subset supports application-oriented research on authentic emotions, while the acted speech subset allows for state-of-the-art emotion categorization comparisons. Fourth, in contrast to existing emotional databases mainly contain only isolated sentences or short dialogs, without taking into account the discourse context, which is known to be an important component. The IEMOCAP corpus was designed from a dialog perspective, eliciting sequential emotions with adequate context. Most interestingly, from an application point of view, this corpus is well-suited for studying the dynamic progression of blended emotions, enabling the detection of when and how a user's blended affective state changes, which can contribute to improving human-computer interactions. This corpus has 12 hours of speech data from ten subjects and is pre-segmented into shortcuts, resulting in a total of 10,039 utterances. Three different evaluators assessed each utterance using the aforementioned nine categorical labels of emotion. It is worth noting that only 22.8% of the utterances in the IEMOCAP corpus are pure examples of a single perception category. Significantly, the remaining 77.2% of utterances present blends of different emotions, which in turn confirms its capability to approach BEES.

B. Task Definition

In contrast to the majority of existing studies in AER, BEES takes a further step by estimating the intensity of each emotion that may concurrently exist in an individual voice stimulus. Different measures have been developed to capture the simultaneity and the subjective experience of emotion-blends [11], [23]. We start by giving a more formal definition of emotional intensity following [24], [25], namely inferring the concurrent experience and intensity of blended emotions by quantifying the frequency with which individuals experience each measured emotion.

Formally, let emoMat be a matrix of experienced and tagged emotional labels for the n^{th} voice stimulus given by three evaluators.

$$\text{emoMat} = \begin{bmatrix} \epsilon^n_{1,1} & \epsilon^n_{1,2} & \cdots & \epsilon^n_{1,\lambda} \\ \epsilon^n_{2,1} & \epsilon^n_{2,2} & \cdots & \epsilon^n_{2,\lambda} \\ \epsilon^n_{3,1} & \epsilon^n_{3,2} & \cdots & \epsilon^n_{3,\lambda} \end{bmatrix} \tag{1}$$

More specifically, each row of the emoMat is correspondingly given by the i^{th} evaluator and λ is the number of emotion classes. Each element of the emoMat is defined as:

$$\epsilon^{\text{n}}_{\text{i},\lambda} = \begin{cases} 1, & \text{if a specific emotion is experienced,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The BEES then calculates the ρ^n_λ as a reference to the emotional vector of the intensity of a specific categorical label λ in the n^{th} voice stimulus as follows:

$$\rho^{\text{n}}_\lambda = \frac{\sum \text{emoMat}\,(:,\lambda)}{\sum_i \sum_\lambda \text{emoMat}_{\text{i},\lambda}} \tag{3}$$

Figure 1 shows an example of the ρ_λ relative to two voice stimuli in the IEMOCAP corpus. The X-axis represents the

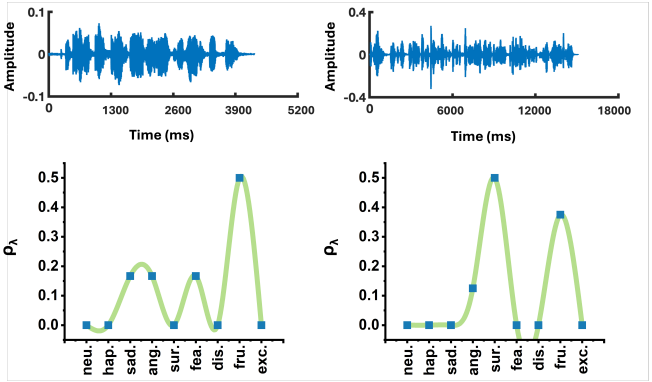

Fig. 1: Two voice stimulus examples with different ρ_λ from the IEMOCAP corpus (left: Ses02M_impro02_F011.wav; right: Ses02F_impro05_F005.wav). Rather than a dominant emotion, human voices often evoke multiple emotions with different perception intensities.

nine kinds of emotions and the Y-axis represents the ρ_λ of each kind of emotion. Note that ρ_λ^n denotes the proportion that λ accounts for in a full emotion distribution of the n^{th} voice stimulus. This differs from the probability of λ being the correct emotion label for the n^{th} voice stimulus. Probability distribution implies that only one emotion label is correct for each sentence, whereas BEES allows for the possibility of multiple emotions co-exist within a single voice.

Specifically, the goal of BEES to identify the component of intensity in terms of each specific emotion λ, ρ_λ^n, can be formulated by computing the optimal parameter θ^* via solving the problem as follows:

$$\theta^* = \text{argmin}_\theta \sum^{N} \text{Dist}(\rho_\lambda^n, \hat{\rho_\lambda^n}) \tag{4}$$

where Dist$(\cdot)$ is the distance function measuring the similarity of the predicted emotion intensity, $\hat{\rho_\lambda^n}$, and the ground truth intensity ρ_λ^n. Additionally, N is the total number of voice stimuli in the IEMOCAP corpus.

III. Baseline Approaches

The overall architecture of BEES is shown in Fig. 2, starting from the human voices and ending with its emotion distribution estimation. The subsequent sections detail the BEES as follows: Section 3.1 introduces the acoustic features that we used to represent voices. Section 3.2 defines a deep neural network-based baseline model to estimate emotion intensity distribution. Section 3.3 provides the metrics to evaluate the BEES performance.

A. Acoustic Features

An important issue in the design of a BEES system is the extraction of suitable features that best reflect different emotions and their intensities and should be robust against other speech diversities of speaking styles, speakers, genders, etc. While deep learning started in this field in the early 2010s, a widespread trend exists to explore the neural representation of acoustics [26], [27]. Within its areas of interest, SSLs have recently gained considerable attention due to their advances in capturing many acoustic characteristics comprehensively and reliably [26], [28]. The importance of the SSLs of speech is increasingly evident in AER, and also many others, like speech

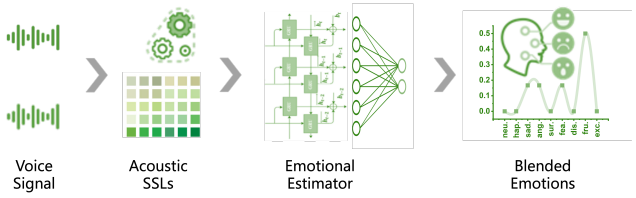

Fig. 2: Illustration of the BEES system. Given a voice evaluated with emotional intensities, our approach represents the voice with a vector matrix using SSLs and then employs a deep neural network-based estimator on it. The mean square error loss is applied to blended emotion estimation.

TABLE I: The BEES results in terms of MCS values to the emoMat references for the overall evaluated emotion intensity obtained by using different SSLs. The top performance is highlighted, and ↑ indicates "the larger the better".

Features	MCS ↑
Data2vec	0.7190
Unispeech	0.6886
HuBERT	0.7087
Wav2vec 2.0	0.6888
WavLM	**0.7303**

recognition [29], speaker diarization [30], gender recognition [31], and more [32], [33]. Five SSLs applied to BEES are thus explored in this work including Wav2vec 2.0[1], HuBERT[2], WavLM[3], Unispeech[4], and Data2vec[5] based on their top performance in well-known benchmarks such as the Speech processing Universal PERformance baseline (SUPERB) [34], and the Holistic Evaluation of Audio Representations [35], as well as recent SER studies [36], [37]. Given that the larger models in general deliver superior performances, achieving state-of-the-art scores on SUPERB tasks [34]. This study therefore approaches MER using large Wav2vec 2.0, HuBERT, WavLM, Unispeech, and Data2vec models, where the last hidden states are transformed into 1024-dimensional vectors for individual speech. For consistent input audio processing, all models are uniformly sampled at 16kHz.

B. Model Configuration

For all SSLs, we employed the same deep neural network-based estimator. We conducted experiments using a hybrid model by combining a bidirectional gated recurrent unit (BiGRU) with a deep neural network. To be more explicit, the BiGRU is utilized to process the acoustic features of SSLs in both forward and backward directions, it could capture contextual information effectively, which is crucial for understanding the dynamic nature of vocal emotion. The bidirectional nature of the BiGRU exhibits robustness to variations in voice stimuli, including differences in speaking rate, accent, and pronunciation [38], [39], enabling it to adapt to diverse speaking styles and linguistic nuances inherent in different speakers. 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 [40], which uses a global average pooling layer instead of a fully connected layer to obtain the SER results. Most notably, this study uses the loss function of mean square error for BEES tasks. Experiments are conducted to estimate blended emotional states via leave-one-speaker-out cross-validation, 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.

[1]https://huggingface.co/facebook/wav2vec2-large
[2]https://huggingface.co/facebook/hubert-large-ls960-ft
[3]https://huggingface.co/microsoft/wavlm-large
[4]https://huggingface.co/microsoft/unispeech-sat-large
[5]https://huggingface.co/facebook/data2vec-audio-large-960h

TABLE II: The BEES results in terms of CCC and MAE values to references for the evaluated emotion intensity considering each specific emotional state in the emoMat obtained by different SSLs. ↑ indicates "the larger the better" and indicates ↓ "the smaller the better".

Features	ang.	dis.	exc.	fea.	fru.	hap.	neu.	sad.	sur.
					CCC ↑				
Data2vec	0.6581	0.0140	0.5465	0.0487	0.4305	0.4575	0.4250	**0.6839**	0.3815
Unispeech	0.5633	0.0304	0.3900	**0.1952**	0.2106	0.2498	0.2933	0.5519	0.2158
HuBERT	0.6592	0.0084	0.5573	0.0044	0.4520	0.4437	0.4350	0.6220	0.0477
Wav2vec 2.0	0.6181	0.0213	0.4631	0.0145	0.3724	0.3473	0.3758	0.6144	0.3166
WavLM	**0.6788**	**0.0682**	**0.5976**	0.0167	**0.5028**	**0.5188**	**0.4764**	0.6651	**0.4327**
					MAE ↓				
Data2vec	0.1130	**0.0060**	0.1163	0.0223	0.2251	**0.0928**	0.1866	**0.0950**	**0.0224**
Unispeech	0.1438	0.0245	0.1413	**0.0157**	0.2382	0.1184	0.2120	0.1111	0.0268
HuBERT	0.1206	0.0133	0.1201	0.0223	0.2108	0.1003	0.1789	0.1027	0.0437
Wav2vec 2.0	0.1307	0.0084	0.1229	0.0216	0.2204	0.1123	0.1976	0.1100	0.0316
WavLM	**0.1039**	0.0070	**0.1142**	0.0178	**0.1954**	0.0980	**0.1770**	0.1059	0.0238

C. Evaluation Metrics

The mean cosine similarity (MCS), concordance correlation coefficient (CCC), and mean absolute error (MAE) between a system's estimations and human evaluations, are calculated as three metrics, in order to evaluate the BEES performance. In particular, the MCS is merely a preferred metric to evaluate the BEES results, in view of the fact that BEES explores the intensities of multiple emotions that co-exist simultaneously as an emotional vector, where smaller MAE and greater CCC measure the agreement between the outputs of the system and the ground truth considering each specific emotional intensity individually might not result in a good BEES performance.

Formally, $\hat{\rho_\lambda^n}$ represents the estimated intensity reference value for a specific emotion λ in the emoMat of the n^{th} voice stimulus from the IEMOCAP corpus, as determined by a BEES system, and the corresponding value of that given by three human estimators is ρ_λ^n. The MCS, CCC, and MAE are accordingly obtained by:

$$\text{MCS} = \frac{1}{N}\sum_{1}^{N} \frac{\sum_\lambda \rho_\lambda^n \hat{\rho_\lambda^n}}{\sqrt{\sum_\lambda \left(\rho_\lambda^n\right)^2}\sqrt{\sum_\lambda \left(\hat{\rho_\lambda^n}\right)^2}} \tag{5}$$

$$\text{CCC} = \frac{2\wp\sigma_{\hat{\rho_\lambda^n}}\sigma_{\rho_\lambda^n}}{\sigma^2_{\hat{\rho_\lambda^n}} + \sigma^2_{\rho_\lambda^n} + \left(\mu_{\hat{\rho_\lambda^n}} - \mu_{\rho_\lambda^n}\right)^2} \tag{6}$$

$$\text{MAE} = \frac{1}{N}\sum_{1}^{N} \mid \rho_\lambda^n - \hat{\rho_\lambda^n} \mid \tag{7}$$

where $\mu_{(\cdot)}$ and $\sigma^2_{(\cdot)}$ are the mean values and variances, respectively, and $\wp$ is the correlation coefficient. Notably, MCS and CCC assign values that trend to 1 for a closer system's estimation to human evaluations; and MAE assigns values that trend to 0 for a better BEES performance of a system's estimations.

IV. Experiment Results and Discussions

In this section, we introduce the BEES baseline using the IEMOCAP corpus and present the main results in Tables I and II. Table I clearly demonstrates that the MCS value consistently turned out to receive a notable gain from the wavLM in comparison with that of other SSLs, providing the highest MCS reaching up to 0.7303. The main reason might be first attributed to the fact that the WavLM utilizes utterance mixing augmentation by adding both interfering speech and noise during the training process. This augmentation strategy significantly contributes to enhancing the model's robustness by exposing it to a more diverse array of inputs. Secondly, given that WavLM extends the HuBERT framework by integrating a gated relative position bias into the Transformer structure, it is poised to introduce advancements in effectively managing relative positions within the sequence that have been shown to play a vital role in conveying emotional information [41]–[43].

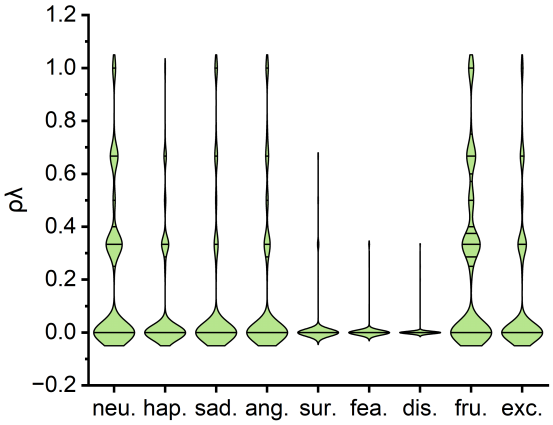

Fig. 3: The ρ_λ distributions for each specific emotion.

Moreover, Table II details the BEES results obtained using different SSLs, with CCC and MAE values provided for each specific emotion, respectively. It highlights that estimating emotional intensities for disgust, fear, and surprise poses notably greater challenges compared to other emotional states. Figure 3 shows a data visualization, illustrating that such difficulty in estimating these three emotions mainly stems from

TABLE III: Comparison results of MCS values on genders (male vs. female) and voice naturalness (acted vs. spontaneous). n.s. indicates that the MCS results do not differ significantly between male and female voices; * indicates that the MCS results differ significantly between acted and spontaneous voices ($p<0.05$).

MCS	Genders$^{n.s.}$		Voice naturalness*	
	male	female	acted	spontaneous
Data2Vec	0.7187	0.7194	0.7380	0.6982
Unispeech	0.6875	0.6898	0.6884	0.6888
HuBert	0.7100	0.7072	0.7293	0.6860
Wav2vec 2.0	0.6932	0.6840	0.6943	0.6828
WavLM	0.7293	0.7314	0.7422	0.7172

their limited distribution of emotional intensity, which results in an unbalanced data problem recognized as one of the main obstacles in the field of machine learning [44].

For further analysis, Table III summarizes two aspects of the BEES results in terms of MCS values, focusing on the impact of genders and voice naturalness. It is important to note that the t-test conducted on all SSLs indicated no statistically significant difference in performance between male and female voices (p=0.56). This observation shows the potential of SSLs in effectively characterizing gender diversity in acoustic emotion analysis, a task known for its complexity in many emotion studies [45], [46]. Additionally, as reasonably expected, the MCS values were significantly higher for the acted voices compared to the spontaneous ones ($p<0.05$), which aligns with findings from prior studies [4], [46]. These results are primarily attributed to the acted voices generally exhibiting fewer clear emotions with higher intensities, while the spontaneous ones tend to be more ambiguous and imprecise, which may be harder to estimate and thus limit their performance.

V. Conclusions

This paper introduces a groundbreaking acoustic task for BEES, addressing the limitations of traditional single-label classification approaches in AER. Our task is well-matched with the general psycho-evolutionary theory of emotion, which posits that human emotions often comprise blends of different emotions with varying intensities. We have offered a straightforward and easily replicable guide for investigating BEES. Future research endeavors could delve into exploring dynamic shifts in BEES within conversational dialogues, providing deeper insights into real-time emotional fluctuations. Moreover, there is promising potential for studying the intricate interplay between emotions and related psychological phenomena, such as stress and depression detection. This proposition not only paves the way for pioneering research in AER but also holds considerable promise for applications in mental health analysis. We hope that BEES will inspire further research and development, leading to powerful emotional agents capable of nuanced emotional understanding and response.

Reference

[1] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[2] P. A. Thoits, "The sociology of emotions," *Annual review of sociology*, vol. 15, no. 1, pp. 317–342, 1989.

[3] L. F. Barrett and J. A. Russell, *The psychological construction of emotion*. Guilford Publications, 2014.

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

[5] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *AI Rev.*, vol. 43, pp. 155–177, 2015.

[6] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[7] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," in *Proc. Interspeech*, 2024, pp. 2024–2350.

[8] J. Tian, D. Hu, X. Shi, *et al.*, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, 2023, pp. 67–73.

[9] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[10] K. R. Scherer, "Analyzing emotion blends," in *Proceedings of the 10th Conference of the International Society for Research on Emotions*, ISRE Publications Würzburg, 1998, pp. 142–148.

[11] R. Berrios, P. Totterdell, and S. Kellett, "Eliciting mixed emotions: A meta-analysis comparing models, types, and measures," *Frontiers in psychology*, vol. 6, p. 428, 2015.

[12] V. Y. Oh and E. M. Tong, "Specificity in the study of mixed emotions: A theoretical framework," *Personality and Social Psychology Review*, vol. 26, no. 4, pp. 283–314, 2022.

[13] A. Israelsson, A. Seiger, and P. Laukka, "Blended emotions can be accurately recognized from dynamic facial and vocal expressions," *J. Nonverbal Behav.*, pp. 1–18, 2023.

[14] P. N. Juslin, P. Laukka, L. Harmat, and M. Ovsiannikow, "Spontaneous vocal expressions from everyday life convey discrete emotions to listeners.," *Emotion*, vol. 21, no. 6, p. 1281, 2021.

[15] E. B. Andrade and J. B. Cohen, "On the consumption of negative feelings," *J. Consumer Res.*, vol. 34, no. 3, pp. 283–300, 2007.

[16] S. H. Hemenover and U. Schimmack, "That's disgusting, but very amusing: Mixed feelings of amusement

and disgust," *Cognition and Emotion*, vol. 21, no. 5, pp. 1102–1113, 2007.

[17] C. C. Bee and R. Madrigal, "Consumer uncertainty: The influence of anticipatory emotions on ambivalence, attitudes, and intentions," *J. Consum. Behav.*, vol. 12, no. 5, pp. 370–381, 2013.

[18] X. Li, Z. Zhang, C. Gan, and Y. Xiang, "Multi-label speech emotion recognition via inter-class difference loss under response residual network," *IEEE TMM*, 2022.

[19] A. Slimi, N. Hafar, M. Zrigui, and H. Nicolas, "Multiple models fusion for multi-label classification in speech emotion recognition systems," *Proc. Comput. Sci.*, vol. 207, pp. 2875–2882, 2022.

[20] X. Geng, "Label distribution learning," *IEEE TKDE*, vol. 28, no. 7, pp. 1734–1748, 2016.

[21] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF conference on CVPR*, 2020, pp. 13 984–13 993.

[22] C. Busso, M. Bulut, C. C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[23] U. Schimmack, "Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive?" *Cognition & Emotion*, vol. 15, no. 1, pp. 81–97, 2001.

[24] K. Oatley and P. N. Johnson-Laird, "The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction," in *Striving and feeling*, Psychology Press, 2014, pp. 363–393.

[25] A. P. McGraw and C. Warren, "Benign violations: Making immoral behavior funny," *Psychol. Sci.*, vol. 21, no. 8, pp. 1141–1149, 2010.

[26] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affect. Comput.*, 2022.

[27] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. on ASLP*, vol. 26, no. 12, pp. 2423–2435, 2018.

[28] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "Using semi-supervised learning for monaural time-domain speech separation with a self-supervised learning-based si-snr estimator," in *Interspeech 2023*. ISCA, 2023.

[29] H. Shi, M. Mimura, and T. Kawahara, "Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[30] V. P. Minotto, C. R. Jung, and B. Lee, "Multimodal multi-channel on-line speaker diarization using sensor fusion through svm," *IEEE TMM*, vol. 17, no. 10, pp. 1694–1705, 2015.

[31] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *MLKE*, vol. 1, no. 1, pp. 492–503, 2019.

[32] K. Fujita, T. Ashihara, H. Kanagawa, T. Moriya, and Y. Ijima, "Zero-shot text-to-speech synthesis conditioned using self-supervised speech representation model," *arXiv preprint arXiv:2304.11976*, 2023.

[33] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," *arXiv preprint arXiv:2205.09812*, 2022.

[34] S. Yang, P. Chi, Y. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[35] J. Turian, J. Shier, H. R. Khan, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 CD Track*, PMLR, 2022, pp. 125–145.

[36] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[37] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 051–11 055.

[38] A. Mutawa, "An end-to-end tacotron model versus pre trained tacotron model for arabic text-to-speech synthesis," *J. of Engineering Research*, 2023.

[39] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, and M. Shafi, "Deepresgru: Residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition," *Knowledge-Based Systems*, vol. 238, p. 107 914, 2022.

[40] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

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

[42] X. Wu, Y. Cao, H. Lu, *et al.*, "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Trans. on ASLP*, vol. 29, pp. 3280–3291, 2021.

[43] X. Li, X. Shi, D. Hu, *et al.*, "Music theory-inspired acoustic representation for speech emotion recognition," *IEEE/ACM Trans. on ASLP*, vol. 31, pp. 2534–2547, 2023.

[44] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *PIAI*, vol. 5, no. 4, pp. 221–232, 2016.

[45] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *PR*, vol. 44, no. 3, pp. 572–587, 2011.

[46] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, 2007.