INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

From Interval to Ordinal: A HMM based Approach for Emotion Label Conversion

Jingyao Wu[1], *Ting Dang*[2], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Nokia Bell Labs, Cambridge, UK

jingyao.wu@unsw.edu.au, ting.dang@nokia-bell-labs.com, v.sethu@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

Ordinal labels along affect dimensions are garnering increasing interest in computation paralinguistics. However, they are rarely obtained directly from raters, and instead typically obtained by conversion from interval labels. Current approaches to such conversion map interval labels to either absolute ordinal labels (AOL) (e.g., low and high), or to relative ordinal labels (ROL) (e.g., one has higher arousal than the other), but never take both into account. This paper presents a novel approach to map time-continuous interval labels to time-continuous ordinal labels. It simultaneously considers both inter-rater ambiguity about where AOLs sit on the interval label scale and the consistency amongst different raters in terms of ROLs. We validate the proposed approach by comparing the converted ordinal labels to original interval labels and the categorical labels for the same speech using the publicly available MSP-Podcast and MSP-Conversation corpora.

Index Terms: Emotion recognition, ordinal labels, label conversion, hidden markov model

1. Introduction

Consequently, speech is one of the most natural forms of human communication and a key modality through which emotions are expressed. Speech emotion recognition plays an important role in human-machine interactions and has received increasing interest in affective computing [1]. Human emotions are complex and can exhibit varying degrees of ambiguity [2]. It has been suggested ordinal emotion representations are better aligned with human perceptions and exhibit less ambiguity compared to other emotion representations [3]. Numerous studies have demonstrated that humans excel at comparing two options rather than assigning absolute labels [4–6]. Thus, a growing body of work has recognised their importance and adopted ordinal regression or preference learning frameworks for emotion recognition [7–10]. More importantly, increasing attention has been drawn to modelling emotions in a time-continuous manner that fulfils the dynamic nature of emotion. This leads to a more interesting challenge of modelling time continuous ordinal labels where more and more works fall into [3]. Nevertheless, there are still great limitations in obtaining such ordinal labels in the first place. This is because collecting primary continuous ordinal labels from human annotators is challenging, thus such emotion datasets are rarely available [3].

It should be noted that two distinct notions of ordinal affect labels exist [11]. Namely, *Absolute Ordinal Labels* (AOLs) which use an ordinal scale to describe affective attributes such as a scale of {low, medium, high} valence; and *Relative Ordinal Labels* (ROLs) which encode pairwise comparisons between instances such as sample *A* has a higher arousal intensity than sample *B*. Both AOLs and ROLs are ordinal but are not equivalent and convey complementary information.

As previously mentioned, challenges inherent in directly collecting ordinal labels have meant most currently available emotion corpora use interval labels and only a very few utilise ordinal labels. Furthermore, even these only provide either AOL [12, 13] or ROL [9, 14], but not both and in all cases the labels are provided per utterance and are not time-varying continuous annotations. Consequently, there is a need to convert interval labels to ordinal labels [3]. Current approaches to conversion from interval labels to AOLs typically assign hard thresholds to partition the interval label space and assign an AOL to each. They only vary in how the thresholds are chosen, which can range from using label cluster boundaries [7,15] to a grid search to obtain a desired trade-off between class balance and inter-rater agreement [11]. Approaches to converting interval labels to ROLs typically involve pairwise comparisons between all pairs of interval labels at different times and retain information about the relationship (greater/lesser) and discarding details of interval label values [9, 16, 17]. None of the current approaches take both *absolute* and *relative* ordinal information into account simultaneously. In this paper we present a novel approach to convert time-continuous interval labels to time-continuous AOLs based on a hidden Markov model that integrates rater specific absolute ordinal state distributions with transition probabilities based on relative ordinal information.

2. From Interval to Ordinal: A State Space Model

Given a time series of arousal/valence ratings, $\boldsymbol{y}^r = y^r_{1:T}$, obtained from rater r over the time interval $t \in [1, T]$, we aim to find a corresponding time series of AOLs, $\boldsymbol{S}^r = S^r_{1:T}$, with each element drawn from a finite set of N possible AOLs $\boldsymbol{\gamma} = \{\gamma_1, \cdots, \gamma_N\}$. i.e., $S^r_t \in \boldsymbol{\gamma}$. For instance, if $N = 3$, $\boldsymbol{\gamma}$ may be {Low, Medium, and High} and the problem can be viewed as that of converting a time-series of numerical arousal/valence ratings to a sequence of Low/Medium/High labels.

Figure 1 (a) shows an example of the original interval labels and the converted AOLs from two raters annotated on arousal for the same speech utterance. It can be seen that the interval scales are significantly different (solid and dash lines). Whereas the AOL sequences (colored regions) are decoded for each individual raters based on their interval labels, such that different regions of interval labels from two raters can map to the same AOL γ_n. This allows the conversion to take into account the heterogeneity in human's perception of emotion.

Additionally, in order to incorporate information about relative changes in ratings that multiple raters agree on, we adopt a hidden Markov model (HMM) to represent the relationship be-

10.21437/Interspeech.2023-2213

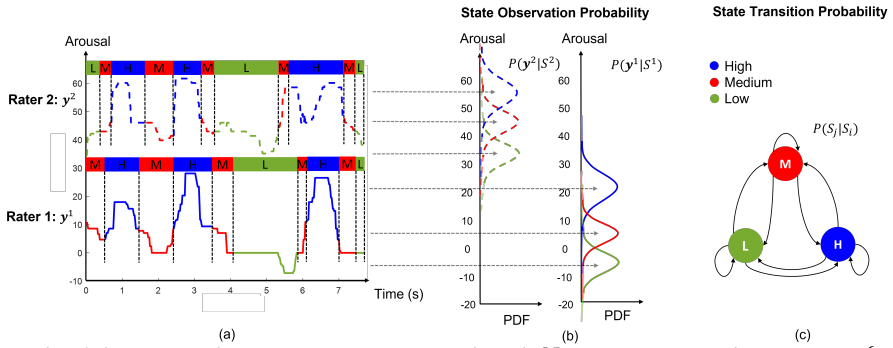

Figure 1: *An example of the proposed AOL conversion approach with $N = 3$, corresponding to states {Low, Medium, High}. (a) A graphical representation of the interval labels and the corresponding converted AOLs from two raters. (b) The state observation probabilities for each rater with each PDF corresponding to one state $P(\boldsymbol{y}|S_i^r)$; (c) A graphical representation of the HMM state transitions among the three states.*

tween interval and absolute ordinal labels that also incorporates information from relative ordinal labels. We choose the number of states of the HMM to be equal to the number of AOL states N, and encode the temporal dynamics of the time-continuous AOLs in the state transitions. Finally, the HMM is individually trained for each rater at each utterance given their individual interval labels, from which it is able to model each rater's annotation characteristics (the range of numerical scores they use corresponding to each ordinal label).

2.1. Hidden Markov Model of Ordinal Labels

The proposed HMM consists of two key components: (i) state observation probabilities, $b_n^r(y) = P(y|\gamma_n, r), \forall n \in [1, N]$, that represents the distribution of interval scores corresponding to each state n of the HMM (one for each ordinal label) for each rate r; and (ii) state transition probabilities, $a_{ij}^r = P(\gamma_j|\gamma_i, r), \forall i, j \in [1, N]$, that describes the probability of switching from one ordinal label to another. The transition probabilities of HMM are initialized to be identical for all raters, as they capture the arousal/valence trends which have been observed to have higher agreement amongst raters [3]. They are further optimized during training for each rater to capture their individual transition characteristics. Henceforth we drop the superscript r when we do not need to distinguish between raters.

We infer the ordinal sequence $\boldsymbol{S} = S_{1:T}$ as:

$$S_t = \arg\max_{\gamma_n \in \boldsymbol{\gamma}} P(S_t = \gamma_n|\theta, y_{1:T}), \quad 1 \leq t \leq T \tag{1}$$

where $\theta = (A, B, \pi)$ denotes the model parameters with $A = \{a_{ij}\}$ indicating the transition probabilities; $B = \{b_i(y)\}$ indicating the set of observation probabilities; and π is the initial state probability $P(S_1 = \gamma_n)$. Finally, we assume that the state observation probabilities are all Gaussian distributions:

$$b_n(y) = P(y|S_t = \gamma_n) = \mathcal{N}(y|\mu_n, \sigma_n) \tag{2}$$

where μ_n and σ_n refer to the mean and standard deviation of the n^{th} Gaussian mixture component respectively.

2.2. HMM Parameter Estimation

Prior to training the HMM, the parameters are initialized to reasonable values in a data driven manner. Specifically, to initialize the state observation probabilities, we first fit an N-mixture Gaussian Mixture Model (GMM) to each rating, $\boldsymbol{y}$:

$$P(\boldsymbol{y}) = \prod_{t=1}^{T} \sum_{n=1}^{N} \omega_n \mathcal{N}(y_t|\mu_n^{(\boldsymbol{y})}, \sigma_n^{(\boldsymbol{y})}) \tag{3}$$

where ω_n indicates the weight of each mixture component.The mixtures are then sorted in order of their means, μ_n, and the observation probabilities of each state of the HMM, $b_n(y)$ is initialised in an increasing order of the sorted mixture components. i.e., the Gaussian mixture component corresponding to the lowest μ_n is set as the initial parameters corresponding to the lowest AOL and so on.

To initialise the transition probabilities, a_{ij}, we take into account the fact that human emotion and, consequently, the affect labels do not change rapidly. Consequently, we expect self-transitions of the states to be fairly high and the initial transition probabilities are assumed to be:

$$a_{ij} = \begin{cases} 1 - \alpha & , i = j \\ \frac{\alpha}{N-1} & , i \neq j \end{cases} \tag{4}$$

where α is a small value, empirically chosen as 0.1.

During optimization, individual GMMs are estimated to each individual rater for initial state observation probability. On the other hand, the initial transition probabilities are set to be the same using Eq. (4) to reflect the higher degree of agreement among raters about relative changes in affect labels. All N initial state distribution, π, are set to be $\frac{1}{N}$. Finally, the Baum-Welch algorithm is used to train the HMM [18], and Viterbi decoding [19] is used to infer the ordinal sequence, $\boldsymbol{S}$.

3. Experimental Settings

To validate the converted AOLs, we compare it with both the original interval labels as well as categorical labels obtained independently to describe the same data. Consequently, databases containing both labels were required, and amongst the publically available datasets, there was only one choice. We selected the MSP-Podcast [13] and MSP-Conversation [20] corpora as they both contain the same recordings but were annotated using the two different label types.

The MSP-Podcast corpus is a large naturalistic speech emotional dataset that is collected from real-life podcast recordings [13]. It contains around 100 hours speaker turns with each segmented into 2.75s to 11s. The speech turns are labelled with categorical labels at each turn by a number of raters varying from 5 to 11. There are more than six emotion categories included in the dataset, however, we only consider the four primary categories that are labelled the most frequently: {Nautral, Happy, Sad, Angry}. In addition, we selected the speaker turns that are labelled with more than 60% agreement among the multiple raters. Consensus categorical labels were selected based on the majority vote.

The MSP-Conversation dataset is a subset of MSP-Podcast dataset [20]. It is labelled with time continuous ratings within

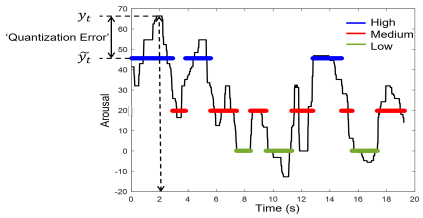

Figure 2: *An example plot of original interval labels $y_{1:T}$ and mean of HMM state $\tilde{y}_{1:T}$, with different colors indicating different states. The error bar indicates the 'quantization error' between them at time t.*

Table 1: *Emotion Change Consensus between the interval labels and converted AOLs computed over different periods (L).*

Consensus Information				
	Arousal		Valence	
L(s)	Baseline ($\widetilde{\boldsymbol{S}}$)	HMM ($\boldsymbol{S}$)	Baseline ($\widetilde{\boldsymbol{S}}$)	HMM ($\boldsymbol{S}$)
1	0.487	**0.545**	0.475	**0.545**
3	0.515	**0.594**	0.512	**0.593**
5	0.562	**0.633**	0.544	**0.630**

the range [-100, 100] on arousal, valence and dominance by 4-11 different raters. Prepossessing is applied to the raw labels following the method in [20] and the resultant labels have a sampling rate of 59 fps. A moving average filter of 1-second window size is applied to smooth the labels. We select the 4 most agreed raters based on the Cronbach alpha as suggested in [20] to achieve a consistent number of raters throughout all utterances. To align the interval labels with corresponding categorical labels for the validation, we only select the speaker turns that are included in both databases. Consequently, we used 2222 turns with each containing 500-1200 samples depending on the lengths of the speaker turns throughout this work.

In all experimental work reported in this paper, the number of states used in the HMM was $N = 3$, leading to the states γ_n – 'low (L)', 'medium (M)' and 'high (H)' for arousal and valence. Thus, three mixtures of GMM was used associating with $\{L, M, H\}$. HMMs were trained using the HMM toolbox [21]. The state initial probability was set to be $\pi_0 = [\frac{1}{3}\,\frac{1}{3}\,\frac{1}{3}]$, and state transition probabilities were initialized using $\alpha = 0.1$.

4. Comparing Ordinal and Interval Labels

In order to compare the converted AOLs to the original interval ratings, we adopt three measures: (i) we estimate how closely the ordinal labels represent the original interval ratings in terms of a 'quantisation error'; (ii) we quantify the consensus in emotion change between the original and converted labels; and (iii) we compare inter-rater agreement before and after conversion. The converted AOLs should achieve higher agreement among the raters compared to the original ratings, while simultaneously retaining pertinent information about emotion level and emotion change. The results are compared to a baseline approach whereby AOLs are obtained as the most likely state, $\widetilde{\boldsymbol{S}}$, based on the state observation probabilities (not making use of the transition probabilities). i.e., $\widetilde{S}_t = \arg\max_{\gamma_n \in \boldsymbol{\gamma}} b_n(y_t)$. This is equivalent to a hard thresholding, the current de-facto approach for mapping interval labels to AOLs.

4.1. Conversion 'Quantisation Error'

When an interval rating time series, $\boldsymbol{y} = y_{1:T}$, is converted to a sequence of ordinal labels, $\boldsymbol{S} = S_{1:T}$, we can view $\boldsymbol{S}$ as a 'quantised' version of $\boldsymbol{y}$ with $\mu_n; n = 1:N$ as the N 'quantisation levels'. This lets us define a 'quantisation error', η_Q, which estimates how well the ordinal labels track the variations in the interval label (refer Figure 2) as:

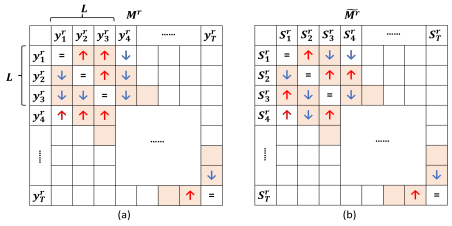

Figure 3: *Example comparison matrices for: (a) original interval labels, M^r; and (b) converted AOLs, $\widetilde{M}^r$. The up and down arrows indicate increase and decrease. The shaded areas indicate the past and future L-second samples being considered for estimating consensus.*

$$\eta_Q = \frac{1}{T}\sqrt{\sum_{t=1}^{T}(y_t - \tilde{y}_t)^2} \tag{5}$$

where $\tilde{y}_t = \mu_n$ when $S_t = \gamma_n$ (noting that $S \in \boldsymbol{\gamma}$).

The averaged η_Q was found to be 4.88±0.7 and 5.17±1.01 for arousal and valence respectively. The errors are reasonably small and are only 2.4% and 2.5% of the full range of [-100,100] on the ratings, suggesting that the converted labels are able to capture major trends in the original ratings while leaving out the small variations. In comparison, the averaged η_Q calculated from the above mentioned baseline, $\widetilde{\boldsymbol{S}}$, was 5.24±0.9 and 6.77±1.21 for arousal and valence respectively. Finally, the separation between the AOLs (low, medium and high) are on average 27.8 and 30.2 for arousal and valence respectively (an example can be found in Figure 2). This is much greater than η_Q suggesting that typical errors are much smaller than the state transitions and unlikely to be misleading.

4.2. Emotion Change Consensus

We estimate the consensus in emotion change between the original interval rating, $\boldsymbol{y}^r = y^r_{1:T}$, from rater r and the converted ordinal label sequence, $\boldsymbol{S}^r = S^r_{1:T}$, by inferring the direction of changes of the label at each frame compared to its neighbouring frames and comparing how often these changes match across $\boldsymbol{y}^r$ and $\boldsymbol{S}^r$.

For both $y^r_{1:T}$ and $S^r_{1:T}$, we carry out pair-wise comparisons across all frames, resulting in the comparison matrices M^r and $\widetilde{M}^r$ respectively, as shown in Figure 3. Specifically, each entry in the matrix M^r is given by:

$$M^r_{pq} = \begin{cases} \uparrow, & y^r_q > y^r_p \\ \downarrow, & y^r_q < y^r_p \\ =, & y^r_q = y^r_p \end{cases} \tag{6}$$

where $p, q \in [1, T]$. The degree of consensus in emotion change, ρ^r_u between $\boldsymbol{S}^r$ and $\boldsymbol{y}^r$ is then estimated as:

$$\rho^r_u = \frac{1}{K}\sum_{pq}\left[M^r_{pq} = \widetilde{M}^r_{pq}\right], \; s.t. \; |q-p| < L \;\&\; p \neq q \tag{7}$$

where $[\cdot]$ denotes the Iverson bracket, and K is the total number of elements being compared (noting that we only consider direction of label change between frames that are at most L apart).

Table 1 reports the average ρ^r_u computed for all raters acorss all utterances for arousal and valence for both the proposed HMM based conversion ($\boldsymbol{S}$) and the baseline ($\widetilde{\boldsymbol{S}}$). We set L as 1sec, 3sec and 5sec, which are the durations that have been shown to capture the temporal dynamics of emotion well [22]. In all cases, the proposed HMM based approach significantly outperforms the hard thresholding based baseline. Further, as L increases, the average ρ^r_u also increases. This indicates that when longer term dynamics are considered there is greater consensus between AOLs and the original interval labels.

Table 2: *Inter-rater agreement of the original interval labels and the converted AOLs computed at different L.*

Inter-rater Agreement						
	Arousal			Valence		
L(s)	Interval	MLE	HMM	Interval	MLE	HMM
1	0.450	0.564	**0.774**	0.454	0.552	**0.796**
3	0.482	0.513	**0.576**	0.491	0.565	**0.593**
5	0.542	0.528	**0.543**	0.525	0.530	**0.480**

4.3. Inter-rater Agreement

We use the matrices M^r and $\widetilde{M}^r$ as introduced in section 4.2 to evaluate inter-rater agreement. For each rater and each utterance, we compute the agreement level in a manner similar to that outlined in [16] for both original interval labels and converted AOLs. For each element within the first L off-diagonals in M^r or $\widetilde{M}^r$, we assign it as an 'agreed element' if more than half of the raters assign the same emotion change direction. Then the agreement ratio is computed as the number of 'agreed entry' divided by the total number of entries in the first L off-diagonals.

Table 2 shows the agreement ratio of AOLs and original labels at different off-diagonal periods for arousal and valence. As expected, the inter-rater agreement with ordinal labels is consistently higher than that with the original interval labels for both arousal and valence [3], and the AOLs obtained via the proposed conversion show greater agreement than those obtained via thresholding (baseline).

5. Comparing Ordinal and Categorical Labels

We also compared the converted AOLs to a set of categorical labels, associated with the same underlying speech, but collected independently of the interval labels. Specifically, since information about change in ordinal labels (or equivalently information captured by ROLs) is encoded in the transition probabilities of the HMM in the proposed approach, we test whether the transition probability matrices exhibit distinct patterns during intervals where emotion changes (identified by change in categorical labels).

5.1. Emotion Change Data

The part of the *utterance* in MSP-conversation that contains a categorical label is referred to as a *turn* in MSP-Podcast. We concatenate consecutive turns (selected as outlined in section 3) into an augmented *augmented segment* which exhibits one of the emotion changes in Table 3. Since we use four distinct categorical emotion labels, the possible types of emotion changes, as well as the number of examples of each available in the dataset, are shown in Table 3. As two types of emotion change (i.e., 'Angry-Sad' and 'Sad-Angry') do not exist in this database, we end up with 14 different augmented segments. Additionally, these segments can be further divided into *Change* and *No-Change* segments based on whether the categorical labels corresponding to the two original segments were the same or distinct.

5.2. Comparing Transition Probabilities

In order to test whether the transition probabilities corresponding to *Change* and *No-Change* segments differ, we gather the set of transition probability matrices corresponding to all the raters and segments within a single *No-Change* augmented segment category and compare it to the set of transition probability matrices corresponding to each *Change* augmented segment category. For e.g., we gather 630 transition probability matrices corresponding to all the 'Neutral-Neutral' augmented segments and compare them to the set of 105 transition probability matrices corresponding to 'Neutral-Happy' augment segments.

Table 3: *Number of the segments in each augmented segment. E.g., the number 105 (row 4 and column 1) indicates there are 105 segments labelled as 'Neutral to Happy'.*

	Happy	Sad	Angry	Neutral
Happy	**91**	2	5	114
Sad	4	**9**	-	22
Angry	3	-	**2**	9
Neutral	105	28	5	**630**

Table 4: *KStest results indicating whether H0 was accepted or rejected, for possible state transitions, and corresponding p values when comparing 'Neutral to Happy' and 'Neutral to Neutral' augmented pair.*

	Low	Medium	High
Low	Accept, p = 0.069	Accept, p = 0.794	Reject, p = 0
Medium	Accept, p = 1	Accept, p = 0.677	Reject, p = 0.0018
High	Reject, p = 0	Accept, p = 0.556	Accept, 0.995

For the comparison, we employ a two-sample Kolmogorov-Smirnov test (KStest). Specifically, each of the transition matrices is comprised of 9-elements (by setting $N = 3$), and we assume each element is drawn from a beta distribution (since they are all probabilities bounded between $[0, 1]$). We then perform the KStest on each element, which accepts or rejects the null hypothesis that the beta distribution from which that element was drawn was identical in the two categories we are comparing (for e.g., 'Neutral-Neutral' vs 'Neutral-Happy').

Table 4 shows the test results as to whether the null hypothesis (H0) was accepted or rejected and its corresponding p values for a comparison between 'Neutral - Neutral' and 'Neutral - Happy' on arousal. It can be seen that H0 is generally accepted under 5% significant level at self-transitions (diagonal) which indicates that the two *Beta* distributions being tested are the same. This is true such that the self-transitions at any augmented segment should always be high enough in order to achieve a smooth AOL sequence, leading to reasonable similar distributions. Whereas on the cross-transitions (off-diagonal), the H0 is rejected at some state transitions. For instance, it rejects H0 at state transition of 'Low $\rightarrow$ High' and 'Medium $\rightarrow$ High'. This aligns with the observation that arousal increases for an emotion change from Neutral to Happy, which should result in different transition probability patterns.

Additionally, we compute the rejection ratio of the cross-state transitions of each comparison, then average across all comparisons (noting that we compare every *No-Change* with every *Change* leading to 40 sets of comparisons). The average rejection ratio was 0.60 and 0.59 for arousal and valence respectively. This suggests that it may be possible to distinguish the *Change* and *No-Change* regions by only looking at the transition probability matrices with up to around 60% accuracy. It also indicates that there are distinct patterns in the transition matrices corresponding to the change in the emotion categories.

6. Conclusion

This paper presents a novel approach to convert time-varying interval affect labels to time-varying ordinal affect labels. The proposed method takes into account both the inter-rater variability in how absolute ordinal notions such as 'Low', 'Medium' and 'High' map to interval scales, as well as the consistency amongst raters in identifying change in emotion. Experimental results on the publicly available MSP-Podcast and MSP-Conversations datasets demonstrate that the converted ordinal labels yield higher inter-rater agreement compared to original interval labels, while retaining pertinent information from the original ratings. Additionally, comparisons of the ordinal labels to categorical labels also reveal that label transitions in the ordinal labels reflect changes in categorical labels.

7. References

[1] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, 2022.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[3] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.

[4] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.

[5] M. Junge and R. Reisenzein, "Indirect scaling methods for testing quantitative emotion theories," *Cognition & Emotion*, vol. 27, no. 7, pp. 1247–1275, 2013.

[6] D. Laming, "The relativity of 'absolute'judgements," *British Journal of Mathematical and Statistical Psychology*, vol. 37, no. 2, pp. 152–183, 1984.

[7] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6494–6498.

[8] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction," *Frontiers in Computer Science*, vol. 3, p. 767767, 2021.

[9] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels," *IEEE Transactions on Affective Computing*, 2023.

[10] O. Ekundayo and S. Viriri, "Multilabel convolution neural network for facial expression recognition and ordinal intensity estimation," *PeerJ Computer Science*, vol. 7, p. e736, 2021.

[11] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "A novel markovian framework for integrating absolute and relative ordinal emotion information," *IEEE Transactions on Affective Computing*, 2022.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[14] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 209–222, 2017.

[15] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[16] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," *Interspeech*, 2018.

[17] G. Zoumpourlis and I. Patras, "Pairwise ranking network for affect recognition," in *9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[18] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[19] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[20] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech*, 2020.

[21] K. Murphy, "Hidden markov model (hmm) toolbox for matlab," *http://www. ai. mit. edu/murphyk/Software/HMM/hmm. html*, 1998.

[22] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method." in *Interspeech*, 2016, pp. 3598–3602.