IDENTIFYING SALIENT SUB-UTTERANCE EMOTION DYNAMICS USING FLEXIBLE UNITS AND ESTIMATES OF AFFECTIVE FLOW

Emily Mower Provost *

University of Michigan, Computer Science and Engineering, Ann Arbor, Michigan, USA

emilykmp@umich.edu

ABSTRACT

Emotion recognition is the process of identifying the affective characteristics of an utterance given either static or dynamic descriptions of its signal content. This requires the use of units, windows over which the emotion variation is quantified. However, the appropriate time scale for these units is still an open question. Traditionally, emotion recognition systems have relied upon units of fixed length, whose variation is then modeled over time. This paper takes the view that emotion is expressed over units of variable length. In this paper, variable-length units are introduced and used to capture the local dynamics of emotion at the sub-utterance scale. The results demonstrate that subsets of these local dynamics are salient with respect to emotion class. These salient units provide insight into the natural variation in emotional speech and can be used in a classification framework to achieve performance comparable to the state-of-the-art. This hints at the existence of building blocks that may underlie natural human emotional communication.

Index Terms— Emotion classification, emotion representation, emotion profile, emotogram, emotion unit

1. INTRODUCTION

Emotion classification systems rely heavily on perceptually evaluated databases. However, the human process of emotion evaluation is still not well understood. When classifying emotion, algorithms generally either process sentence-level data or use data windowed over fixed units in time. However, it is not clear that humans, our ground truth, process emotion in this manner. Therefore, these fixed time units may not be appropriate for modeling affective data. This paper discusses methods to: (1) identify variable-length units that capture salient information relating to local dynamics in emotional speech and (2) understand how to use sub-utterance dynamics to identify utterance-level emotion labels.

The importance of local cues has been well studied in human perception, for example [1, 2], which highlight the interaction between local and global cues. The current work investigates the benefit of estimating salient local (sub-utterance) emotion dynamics via variable-length units. For example, consider a portion of the utterance that displays increasing evidence of anger over multiple frames. A unit that can capture this trend may be a stronger indication of anger than any individual frame. The goal is to capture these patterns and to use them in classification to understand how emotion change influences perception.

Recent work in emotion and behavior modeling has focused on methods to approximate the human evaluation process when classifying data, specifically the interaction between context and salience [3]. The work of Lee et al. [4] approximated humans' ability to both identify salient regions of interest and to integrate cues over time. They found that the utility of these two methods depended on the type of behavior being modeled. The work of Wöllmer et al. [5] and Metallinou et al. [6] have closely examined how context can be used in the modeling process. Context has also been studied in dialogue systems [7]. Further, variations of dynamics have been attributed to changes in emotion perception, notably in the music domain [8, 9]. In our previous work, we looked at modeling the intra-utterance dynamics of an utterance using Hidden Markov Models (HMM) [10]. We found that this tracking could effectively capture the utterance-level label and that the choice of unit size impacted accuracy (also demonstrated in [11]). However, this dynamic assessment could not provide insight into the local building blocks of emotional speech due to the restrictions of state-based modeling. The current work investigates alternative methods to approximate the dynamic nature of emotion by allowing for variable-length units, identifies emotionally salient regions using these units, and presents methods to classify the affective label of an utterance by accumulating local salient evidence.

This problem is approached by first estimating the short-time affective content of the data, creating an n-dimensional characterization where n is the number of affective cues detected. These n-dimensional estimates are aggregated into a set, called an emotogram. The emotogram describes how these cues ebb and flow dynamically over an utterance [12]. The emotograms are automatically segmented into variable-length n-dimensional sub-trajectories. The sub-trajectories are clustered to create a discrete set of **Emotion Units** (EU). The salience of each EU is calculated. The emotional relevance of EU saliency is validated in a classification framework.

This framework is a novel approach to learning variable-length units and their relevance to emotion perception. It uses trajectory partitioning [13] to identify variable-length emotion units and salience detection [14] to natively identify the salience of the estimated affective flow. The results show that EUs contain different levels of emotion salience, demonstrating that certain EUs are more strongly associated with particular emotion classes (e.g., a rise in anger strongly suggests the class of "angry"). The classification results demonstrate that this method can achieve comparable results to the state of the art [15]. These findings suggest that there may exist basic building blocks that underlie expressions of emotion.

2. DESCRIPTION OF DATA

The data used in this study are from the USC Interactive Emotional Motion Capture Database (IEMOCAP). The database was collected using mixed-gender pairs of actors performing from scripts and improvised scenarios. The database contains approximately 10 hours of data recorded using audio, video, and motion-capture. In this study the data are restricted to utterances that contain both audio and

*This work is supported by the National Science Foundation (NSF RI 1217183)

motion-capture (approximately five hours due to recording conditions). The data were evaluated using categorical labels (i.e., angry, happy, neutral, sad, frustrated, excited – merged with happy, disgust, fear, surprise, other) and dimensional labels. In this study only categorical labels were used. The evaluator agreement over the categorical labels was 0.40 (given majority consensus). After merging the classes of happiness and excitement the agreement increased to 0.48. The evaluation ambiguity highlights the emotional subtlety of the dataset. The utterance lengths vary from 0.5 seconds to 33.25 seconds. This variation in length provides a great opportunity to understand how evaluators make assessments over utterances of different lengths. The utterances used in this study had majority ground-truth labels of: angry (584), happy (1,153), neutral (515), and sad (571). There were a total of 2,823 utterances.

The audio features used in this database include: pitch, energy, and Mel Filterbanks, extracted using Praat [16]; features demonstrated to be very effective in this domain [12, 17]. The statistics extracted from these features include mean, standard deviation, lower, upper, and quantile range. The statistics for pitch were extracted only over the non-zero portions of the signal. Initial and trailing silence were approximated and removed by discounting portions of the utterances before the first non-zero pitch value and after the last non-zero pitch value. The motion-capture data includes the x-y-z positions of 53 facial markers. The motion-capture features are based on Facial Animation Parameters (FAP, discussed in [18]). FAPs describe the distances between points on the face. The specific features are described in more detail in [19].

The original feature set consists of 685 features. This feature set size was reduced using the Wilcoxon test, a nonparametric test that compares the centers of two populations (implemented using rankfeatures in Matlab). Emotion-specific features sets were selected in a self vs. other paradigm (e.g., the features that best separate anger from not anger). This resulted in four emotion-specific feature sets to be used in a binary self vs. other classification framework. The number of features retained was determined using a parameter sweep over the training data (Section 3). The number of features varied across folds, but was either 180 (3/10 speakers) or 200 features.

3. METHODS

This section describes methods to (1) estimate emotion flow via emotograms, described in [10, 12, 19], (2) estimate variable-length units that capture salient changes in affective flow, and (3) detect emotion class using the salience of these estimated units. All algorithms discussed in this paper use leave-one-subject-out cross-validation. The training data (nine speakers) are entirely subject-disjoint from the test data for each fold. The parameters for each test speaker are selected using leave-one-subject-out cross-validation over the nine training speakers; resulting in nine parameter sets. The final parameter set is the median of the nine best parameter sets. This parameter set is applied to the unseen test data.

Sentence-level emotion dynamics are captured via emotograms [10, 12]. The emotogram for an utterance is the set of short-time affective estimates calculated over windowed portions of that utterance. These short-time estimates are called Emotion Profiles (EP) [19]. Each EP describes the presence or absence of a set of emotional cues over a windowed portion of the utterance. The idea behind the emotogram methodology is that emotion can be better understood if we understand how high-level affective cues ebb and flow over an utterance. More specifically, EPs are n-dimensional vectors containing the classifier-derived confidence, C, in the presence or absence of each emotion in the set of emotional cues. The EP for utterance i can be described as,

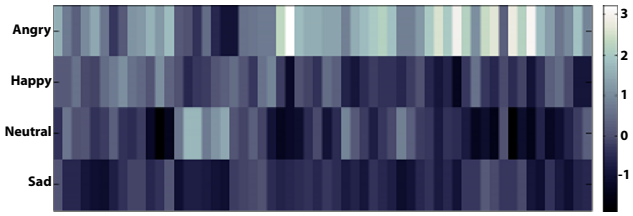

Fig. 1. The emotogram for an angry utterance. The white represents confidence in the presence of an emotion while black represents confidence in the absence of that emotion. The vertical slices are the estimates of emotion at each 0.25 second window. The dynamics of the utterance can be viewed by observing how the estimated presence and absence of the emotional cues ebb and flow over time.

$EP_i = \{C_k\}, k \in \{angry, happy, neutral, sad\}$, for a four-dimensional profile (higher-dimensional extensions have been explored [20]). EPs are estimated using n self vs. other binary Support Vector Machine (SVM) classifiers (here, $n = 4$). The SVM classifiers use radial basis function (RBF) kernels with $\sigma = 9$, selected across all speakers using the parameter sweep. Figure 1 shows an emotogram for an angry utterance. The vertical slices in the figure are EP estimates. Each vertical slice of the emotogram describes the presence (white) or absence (black) of each of the affective cues. All slices are normalized (z-normalization) on a per-speaker basis. In this paper, the emotograms are composed of EPs calculated using a sliding window of 0.25 seconds (with half a window overlap). This results in a set of four **Emotion Trajectories** for each utterance. An emotion trajectory (or "trajectory") is an estimate of how each emotion cue varies over the utterance (e.g., anger variation can be seen in the top line of Figure 1).

3.1. Trajectory Segmentation

The four emotion trajectories provide a dynamic description of the nature of estimated emotion flow over an utterance. These dynamics are used to identify salient regions of the data. These salient regions can then be used for classification or to build a better understanding of the dynamics of emotion expression. Ultimately, it is not clear that the basic building blocks of the emotogram, the fixed window estimates of emotion content, provide the only available unit type for saliency modeling. For example, consider pitch modeling. Statistics of pitch are commonly used features in emotion classification. However, the pitch contour, or how pitch changes over time, has also been shown to be an effective feature for emotion classification [21,22]. In this work, salient static emotogram slices (an analog to the pitch statistics) are compared to salient emotogram dynamics (an analog to the pitch contour). Characteristic emotion dynamics may evolve over different time scales. Consequently, capturing accurate emotion dynamics may require the identification of short variable-length regions of constant emotion change. These variable-length regions (units) are identified using TRACLUS, a trajectory segmentation method presented in [13]. The goal of this work is to uncover these natural, perceptually meaningful, variable-length units to develop an understanding of the dynamics of emotion expression.

The trajectory segmentation is based on minimum description length (MDL), introduced in [23]. MDL attempts to balance two competing constraints: (1) $L(H)$ – the length, in bits, of the description of the hypothesis, H and (2) $L(D|H)$ – the length, in bits, of the description of the data when encoded with the help of the hypothesis. The goal is to find the best hypothesis, H, that will explain the data, D, while minimizing the sum of $L(H)$ and $L(D|H)$. The partitioning is decided using two values, MDL_{par} and MDL_{nopar},

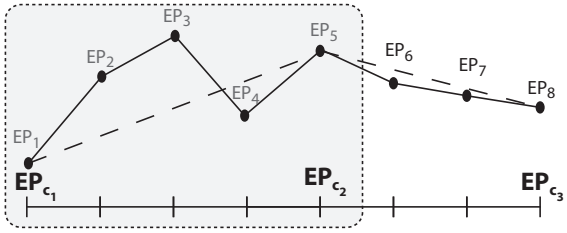

Fig. 2. This figure demonstrates the process of segmenting a trajectory. The original five points are: $\{EP_i\}, 1 \leq i \leq 8$. The question is whether to retain the original five point description (black line) or to form a new line with a start/end points of $EP_{c,1}$ and $EP_{c,2}$ (dotted line). The dotted line trajectory is an approximation of the original black trajectory.

the cost of partitioning and the cost of not partitioning, respectively. MDL_{par} is the sum of $L(H)$ and $L(D|H)$. MDL_{nopar} is merely $L(H)$. If the cost of partitioning the data is greater than the cost of not partitioning the data, then the previous data point is assigned as a characteristic point (CP). The goal is to reduce the size of the trajectory from the number of points in the original trajectory to the set of CPs, $|CP| < |Trajectory|$.

For example, consider an utterance whose emotogram contains eight EPs ($EP_1 ... EP_8$, see Figure 2). During partitioning, the algorithm must decide if the original trajectory (black lines in the figure) specified by the first five EPs should be retained or if instead a simpler line (dotted) could be used as an approximation to this emotional trajectory. The approximated dotted-line trajectories form the sub-trajectories that are discussed throughout the remainder of this paper. In the example, the algorithm returns that the CP locations (c_j) are at window 1, 5, and 8 (EP_{c_1}, etc.). Thus, there are two new sub-trajectories: (1) EP_{c_1} to EP_{c_2} ($c_1 = 1$, $c_2 = 5$) and (2) EP_{c_2} to EP_{c_3} ($c_2 = 5$, $c_3 = 8$). These sub-trajectories capture change in the estimated presence/absence of the emotional cues. Thus, these sub-trajectories are an estimate of local emotion dynamics.

The data, D, is the original segment of the trajectory (e.g., black line: EP_1 to EP_2 in Figure 2). H refers to the proposed sub-trajectory (e.g., dotted line: EP_{c_1} to EP_{c_2} or unchanged black line). The components of the MDL assessment, $L(H)$ and $L(D|H)$ are calculated as in Equations 1 and 2. In the case of MDL_{nopar}, $L(H)$ is the sum of the length of each component of the black trajectory (as seen in Equation 1). In the case of MDL_{par}, $L(H)$ is the length of the new single component of the trajectory (dotted line). $L(D|H)$ describes the aggregated distance between each of the original black line trajectory components and the proposed dotted line trajectory segmentation (e.g., distance between EP_2EP_3 and $EP_{c_1}EP_{c_2}$). The distance is calculated using angular and perpendicular distances. Additional details can be found in [13]. The result of this process is a set of segmented trajectories for each emotogram.

$$L(H) = \sum_{k=c_j}^{c_{j+1}-1} log_2[len(EP_k EP_{k+1})] \quad (1)$$

$$L(D|H) = \sum_{k=c_j}^{c_{j+1}-1} \{log_2[d_\perp(EP_{c_j} EP_{c_{j+1}}, EP_k EP_{k+1}) + log_2[d_\theta(EP_{c_j} EP_{c_{j+1}}, EP_k EP_{k+1})]\} \quad (2)$$

3.2. Salience Modeling

The goal of this work is to identify EUs and to demonstrate that they contain emotionally salient detail. This requires a discrete set of EUs. Discrete labels were assigned to each sub-trajectory using hierarchical clustering (Matlab, PRTools [24]) rather than the trajectory clustering methods of [13]. Hierarchical clustering is a bottom-up approach that merges the two most similar clusters at each stage. The algorithm terminates either with a single cluster or at a pre-specified number of clusters. In this paper the number of clusters was chosen using the parameter sweep method (Section 3). The number of clusters varied from 100 to 200 across the folds. The number of sub-trajectories in the training data were down-sampled by a factor of 100 to mitigate computational complexity (43,252 units to 433 units). Thus, hierarchical clustering can be seen as seeding the initial clusters. The features for clustering included the parallel, perpendicular, and angular distances between segments, demonstrated effective in [13]. The remaining training data and held out test data were assigned to the cluster of the closest training data point.

The emotional relevance of the EUs was assessed using a method proposed by [14] for modeling the affective salience of words in call center databases. The authors identified words associated with positive and negative affective classes. Here, the "words" are the variable-length clustered EUs. Salience was calculated in a self-vs. other paradigm (in the following descriptions, the class of anger will be discussed, the same strategy holds for each emotion class). The emotion-specific salience for any EU, $sal(EU_{i,k})$ is the product of the conditional probability describing the presence of the unit, EU_i, given emotion class k: $p(EU_i|e_k), k \in \{angry, not\ \ angry\}$, and the mutual information, $log(p(EU_i|e_k)/p(e_k))$. The general salience of each EU, $gsal$, is the sum of the emotion-specific saliences, $gsal(EU_i) = \sum_k sal(EU_{i,k})$. EUs with a $gsal$ less than a specified threshold were treated as filler and were discarded. The threshold was set as a quantile of $gsal$, determined using the parameter sweep method; it varied between 0.1 and 0.9 quantiles.

The utterances were classified using accumulated salience, the amount of emotion-specific evidence an evaluator is estimated to receive while observing an utterance. It is the sum of the emotion-specific saliences for EUs with $gsal$ greater than the threshold. This results in a four-dimensional saliency estimate for each test utterance. The class label is assigned based on the highest accumulated salience. The EU saliency model will be referred to as "EU-Sal."

3.3. Alternative Models

The first model uses the saliency modeling described in Section 3 applied to quantized and coded versions of the individual EP slices (e.g., $\{EP_t\}_{t=0}^{t=T_N}$) of the original unsegmented emotogram. This can be seen as salient evidence spotting and accumulation. The four-dimensional EP slices are quantized into bins using the quantiles at 0.2 and 0.8 and the resulting binned slices are coded as 1 – 81 based on bin value (e.g., mid-angry, no-happy/neutral/sad is quantized into bins: 2111 and coded as: code 28) [12]. This produces the one-dimensional emotograms needed for saliency detection. These descriptors are referred to as raw units (RU). The RU saliency model will be referred to as "RU-Sal." The RU-Sal model assumes that windowed short-term evidence provides important cues regarding the high-level label of an utterance (e.g., strong evidence for the presence of anger over a particular window).

The second model uses Hidden Markov Models (HMM) to capture the dynamics of the four-dimensional EP slices via a three-state model with left-to-right topology, described in [10]. This strategy assumes that utterance-level dynamics contribute to overall emotion

Sent. Length	EU-Sal UW	EU-Sal W	RU-Sal UW	RU-Sal W	HMM UW	HMM W
6 - 33.25	71.41	71.06	69.95	71.37	70.64	71.80
3 - 6	65.10	66.12	65.00	66.40	63.94	67.42
1.5 - 3	59.34	60.61	57.68	59.60	60.71	63.32
0.5 - 1.5	59.01	60.36	58.99	59.91	63.38	64.58
Overall	63.80	64.54	62.34	64.16	65.04	66.17

Table 1. Classification results for the presented systems (in percentages). UW Acc. stands for unweighted accuracy (an average of the four emotion class accuracies). W stands for weighted accuracy.

perception and that users integrate all emotional information when making an assessment. Comparable performance between the EU-Sal and HMM methods would suggest that certain local dynamics (in addition to utterance-level global dynamics) are correlated with class label. In [10] varied window sizes were analyzed, in the current work, the window size is restricted to 0.25 seconds to permit HMM modeling across all sentence lengths (0.5 – 33.25 seconds).

3.4. Model Comparison

The RU-Sal model is predicated on the hypothesis that salient local evidence can be accumulated over the course of each test utterance to form an estimate of emotion class. The EU-Sal model asserts that salient local dynamics also contain emotionally relevant cues. The HMM model asserts that global dynamics can be used for classification. These assertions are tested across utterances of different lengths to understand the trade-offs between local and global affective cues. The EU- and RU-Sal models are expected to perform well on long utterances as these utterances often contain natural emotion fluctuations over their span. Evaluators observe these fluctuations and produce an assessment of the high-level content of these clips. The EU-Sal and RU-Sal methods emulate this, identifying salient dynamics and salient evidence over the course of an estimate. Therefore, even given noisy estimates of the affective content, the expectation is that it is possible to identify at least a subset of the salient evidence. It is expected that as the length of an utterance decreases the availability of numerous noise-free estimates also decreases and consequently the opportunity to recognize salient evidence and dynamics similarly decreases. The HMM experiments are expected to perform more accurately on the shorter utterances because the unavailability of precise evidence will be mitigated by the overall affective dynamics of the utterance.

4. RESULTS AND DISCUSSION

The results include weighted and unweighted accuracy calculated over each fold. Unweighted accuracy is the average of the four emotion-specific recall measures for each fold. Weighted accuracy is the fraction of the number of utterances classified correctly over the total number of utterances for each fold. The presented accuracies are the average over all ten folds.

The results demonstrate that the overall accuracies across the three models are similar. Previous research demonstrated the efficacy of global dynamics for emotion classification [10]. The current work suggests that local dynamics can also be effectively used for emotion classification. For utterances greater than six seconds in length, EU-Sal is the most effective technique (UW: 71.41), followed by HMM (UW: 70.64) then RU-Sal (UW: 69.95). For utterance lengths of 3-6 seconds, both the RU-Sal and EU-Sal methods outperform the HMM (UW), 65.10, 65.00 and 63.94, respectively. For utterance lengths below three seconds HMMs consistently outperform the RU-Sal and EU-Sal models.

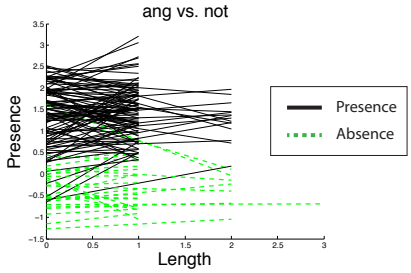

Fig. 3. A subset of the salient segments for presence/absence of anger. Only the angry dimension is shown.

In addition to classification, EU-Sal provides a method to understand the local dynamics that contribute to the perception of a particular class. In the EU-Sal method, every four-dimensional emotogram sub-trajectory is assigned to a cluster. These sub-trajectories can be visualized to gain insight into the types of local dynamics that are salient. Figure 3 presents the general salience of a random subset of emotionally salient trajectories for the differentiation of angry and not angry. The figure shows only the angry EP dimension. The black segments are salient with respect to the class "angry." The green dotted segments are salient with respect to the class "not angry." The salient segments for angry are generally those that provide strong evidence for anger. Conversely, the salient segments for not angry are generally those that describe either low evidence or a decrease in evidence for the class of anger.

In the final analysis, the salience of the variable-length units is discussed. Figure 4 displays the proportion of EUs of a given length (y-axis) that are salient to a particular degree (x-axis). The figure demonstrates that the salient segments of any length are a fraction of the total number of EUs of that length. This finding is reasonable given the thresholding that governs whether or not a segment is considered salient. However, the figure shows that all of the unit lengths contain salient members, suggesting that the variable-length units are important for describing local emotion dynamics.

5. CONCLUSIONS

This paper presents a novel method for natural variable-length unit detection in affective communication. The results demonstrate that variable-length units can effectively capture local dynamics and can be used in a classification framework to achieve results comparable to the state-of-the-art on the IEMOCAP database (62.42% [15]). The results also show that these units can be used to interpret the emotional dynamics of affective utterances. Future work includes investigating the presence of patterns that guide the ordering of these units to gain insight into the structure that underlies emotional speech.

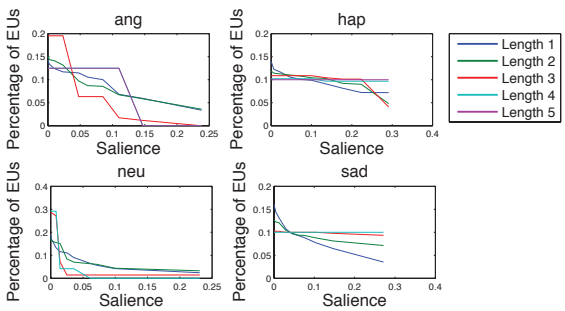

Fig. 4. The relationship between segment length and salience across the emotion classes. The y-axis represents the proportion of EUs of a particular length that have the salience described by the x-axis.

6. REFERENCES

[1] L.D. Sanders and D. Poeppel, "Local and global auditory processing: behavioral and erp evidence," *Neuropsychologia*, vol. 45, no. 6, pp. 1172–1186, 2007.

[2] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.

[3] J. Gibson, A. Katsamanis, M.P. Black, and S. Narayanan, "Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines," in *Interspeech*, Florence, Italy, Aug. 2011.

[4] C.-C. Lee, A. Katsamanis, P.G. Georgiou, and S.S. Narayanan, "Based on isolated saliency or causal integration? toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test," in *Interspeech*, Portland, OR, Sept. 2012.

[5] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Interspeech*, Makuhari, Japan, Sept. 2010, pp. 2362–2365.

[6] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[7] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Interspeech*, Sept. 2005, pp. 1845–1848.

[8] S.B. Kamenetsky, D.S. Hill, and S.E. Trehub, "Effect of tempo and dynamics on the perception of emotion in music," *Psychology of Music*, vol. 25, no. 2, pp. 149–160, 1997.

[9] E.M. Schmidt and Y.E. Kim, "Modeling musical emotion dynamics with conditional random fields," *ISMIR, Miami, FL*, 2011.

[10] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2372–2375.

[11] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm," in *Interspeech*, Makuhari, Japan, Sept. 2010, pp. 801–804.

[12] E. Mower Provost and S. Narayanan, "Simplifying emotion classification through emotion distillationn," in *Proceedings of APSIPA Annual Summit and Conference*, Hollywood, CA, Dec. 2012.

[13] J.G. Lee, J. Han, and K.Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 593–604.

[14] C.M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–302, 2005.

[15] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, March 2010.

[16] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, Sept. 2009, pp. 25–29.

[18] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie, "Emotion Recognition and Synthesis Based on MPEG-4 FAP's," in *MPEG-4 Facial Animation: The Standard, Implementation, and Applications*, I. S. Pandzic and R. Forchheimer, Eds., chapter 9, pp. 141–167. John Wiley & Sons, Ltd., 2002.

[19] E. Mower, M. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[20] E. Mower, K. J. Han, S. Lee, and S. Narayanan, "A cluster-profile representation of emotion using agglomerative hierarchical clustering," in *Interspeech*, Makuhari, Japan, Sept. 2010.

[21] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2003, vol. 2, pp. II–1.

[22] J Vroomen, R. Collier, and S. Mozziconacci, "Duration and intonation in emotional speech," in *Proceedings of the Third European Conference on Speech Communication and Technology, Berlin*, 1993, pp. 577–580.

[23] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[24] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. Tax, and S. Verzakov, "Prtools 4.1," *A Matlab Toolbox for Pattern Recognition, Software and Documentation downloaded May*, 2010.