EMOTION CLASSIFICATION VIA UTTERANCE-LEVEL DYNAMICS: A PATTERN-BASED APPROACH TO CHARACTERIZING AFFECTIVE EXPRESSIONS

Yelin Kim and Emily Mower Provost *

University of Michigan
Electrical Engineering and Computer Science, Ann Arbor, Michigan, USA
yelinkim@umich.edu, emilykmp@umich.edu

ABSTRACT

Human emotion changes continuously and sequentially. This results in dynamics intrinsic to affective communication. One of the goals of automatic emotion recognition research is to computationally represent and analyze these dynamic patterns. In this work, we focus on the global utterance-level dynamics. We are motivated by the hypothesis that global dynamics have emotion-specific variations that can be used to differentiate between emotion classes. Consequently, classification systems that focus on these patterns will be able to make accurate emotional assessments. We quantitatively represent emotion flow within an utterance by estimating short-time affective characteristics. We compare time-series estimates of these characteristics using Dynamic Time Warping, a time-series similarity measure. We demonstrate that this similarity can effectively recognize the affective label of the utterance. The similarity-based pattern modeling outperforms both a feature-based baseline and static modeling. It also provides insight into typical high-level patterns of emotion. We visualize these dynamic patterns and the similarities between the patterns to gain insight into the nature of emotion expression.

Index Terms— emotion classification, emotion dynamics, emotion structure, multimodal, dynamic time warping, dynamic pattern

1. INTRODUCTION

Interest in automatic emotion recognition has grown rapidly in the past few decades. This growth has fueled the development of quantitative models of human emotion dynamics, which augment our interpretation and understanding of these complex expressions. A proper understanding of the dynamic nature of emotion will lead to modeling advancements and a greater understanding of the structure that underlies our affective communication. One of the most common methods for assessing global emotion dynamics is using Hidden Markov Models (HMMs). This technique gained popularity in the speech recognition community and has been effectively used in the emotion recognition community as well. However, in this work, we take a different approach and focus on methods that will provide interpretable descriptions of emotion dynamics. We quantify how emotion flows over an utterance and demonstrate how patterns of this flow can effectively be used to predict an emotion state.

Our work is motivated by *process-oriented* research in Psychology. This approach uses statistical models to forecast or predict psychological behaviors of a human over his/her life span [1–4]. This approach provides a framework to assess human-centered fluctuation. Human emotion is, by its very nature, a variant signal, even over very short time intervals (with respect to the life span of an individual). We hypothesize that certain dynamic patterns may underlie emotional communication. Our goal is to identify these patterns, which we call "flow patterns," and use them in an emotion classification framework. We further hypothesize that the salient characteristics of these patterns are the long-term utterance-level dynamics rather than the short-term fluctuations. We expect to see common patterns repeating over utterances of the same emotion class. We propose a simple quantitative method to model the flow patterns and demonstrate how these patterns of estimated emotion dynamics furthers our understanding of human emotion expression. We estimate emotion flow by extracting features related to emotion dynamics. The features are sequential short-term estimates of emotion states extracted using methods introduced in [5,6]. Each estimate describes the utterance in terms of blends of emotional cues. Our previous work demonstrated that sequential emotion estimates can be used to classify and identify affective states in a dynamic classification setting [6]. In this work we present an emotion modeling technique that leverages the intra-utterance flow patterns to capture the emotional similarity between utterances. This method natively provides insights into the flow patterns and their relationship to emotion state.

There is a large body of work in tracking feature-level emotion structure, including HMMs and Bidirectional Long Short-Term Memory (BLSTM) systems. The BLSTM models are neural networks with memory blocks that can capture variable amount of context [7, 8]. These models are effective in capturing long-range context. However, this context is firmly tied to the multiplicative gate units and may be difficult to interpret [9]. In these methods, and commonly within the community, the common practice for modeling emotion dynamics considers the feature-level fluctuations of the signal [10, 11]. We have demonstrated that short-term estimates of affective flow could also be modeled dynamically using HMMs. This suggested that emotion has definable structure [6]. However, our understanding of these underlying dynamics was restricted by the limitations of a finite state space [6]. In this paper we provide a framework for dynamic modeling of emotion that provides interpretable descriptions of emotion expressions by explicitly focusing on utterance-level dynamics.

Our proposed method captures emotional similarity by estimating time-series similarity between flow patterns of different utterances. This allows us to explicitly take longer-range temporal dependencies into account because our method is focused on variation over the entire utterance (rather than frame level change). We first estimate short-term emotion content over small time windows for each utterance, which approximates emotion dynamics. We calculate the similarity between these estimated dynamics using Dynamic Time Warping (DTW). Unlike HMM, DTW does not make any sta-

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

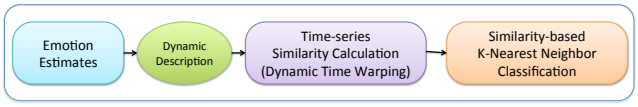

Fig. 1. *Illustration of the proposed method*

tistical assumptions about the intrinsic model. Instead, it directly computes the flow pattern similarity between the unknown utterance and known time-series data [12]. We use this DTW similarity measure in an automatic emotion classification system (Figure 1).

The novelty of this work is in its focus on interpretable utterance-level dynamic modeling, which furthers our understanding of the structure underlying emotional utterances. The results demonstrate that this modeling is effective for identifying emotion state. The maximal accuracy of flow pattern modeling of estimated emotion in DTW similarity-based classification system is 64.40% (unweighted accuracy). This accuracy is greater than that of a baseline model that captures the flow patterns at the feature level as well as a static model, 64.02% and 61.20%, respectively. Further it performs comparably to the state-of-the-art results on a different subset of the same database [13]. This suggests that flow pattern of temporal emotion dynamics provide useful information for emotion recognition.

2. DATA

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [14]. This database contains acted and improvised interaction scenarios between five pairs of actors (one male-one female). The data include audio, video, and motion capture cues with over 12-hours of affective expressions. The data are segmented into over 10,000 utterances about half of which have motion capture information. The categorical ground truth for each utterance was annotated by at least three human evaluators. In this work we use utterances with a ground truth from the set: *Angry, Happy, Neutral, Sad*. There are three different expression types that are used in our work: prototypical, non-prototypical and combined data. Prototypical data are utterances with total evaluator agreement, non-prototypical data have only majority vote agreement, and the combined data include both prototypical and non-prototypical data. There are 284 angry, 707 happy, 123 neutral, and 316 sad prototypical utterances (1430 utterances in total) and 316 angry, 498 happy, 455 neutral, and 319 sad non-prototypical utterances (1588 utterances in total).

In this work we use both audio and motion-capture features. The audio features include pitch, intensity and Mel Filterbank coefficients. The motion-capture features are based on Facial Animation Parameters. Our input features are statistical functionals of the raw features. They include: mean, variance, lower quantile, upper quantile, and quantile range. The initial feature set contained 685 features, where 145 are audio and 540 are video features. This initial feature was reduced to 180 features as in [6] using Principal Feature Analysis (PFA) [15]. PFA is a variant of Principal Component Analysis (PCA). It projects the input data into the PCA space and clusters the data in this space using k-means. It returns the features closest to the center of each cluster. This ensures that the final features are features in the original space and that a target level of variance in the dataset is retained.

3. EMOTION ESTIMATION

3.1. Emotion Profile (EP)

The short-term affective estimates are made using the Emotion Profiles (EPs) framework. EPs were introduced and demonstrated to be

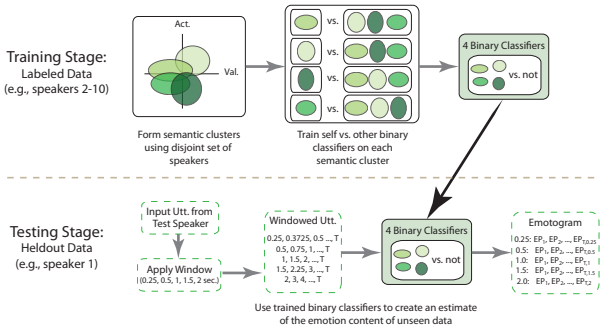

Fig. 2. *Emotion Profiles (EPs) and Emotograms generation proposed and described in the previous work [13]*

effective for emotion recognition tasks in [5, 6, 13]. EPs describe the emotion content of an utterance by capturing the subtle blends of emotional cues present in that utterance. EPs estimate the degree of confidence in the presence or absence of each of these cues, forming an n-dimensional estimate of affective content. In this paper, we use utterances in the label set: *Angry, Happy, Neutral, Sad*. Thus, the EP for an utterance is a four-dimensional vector describing the confidence, c, in the presence of each emotion from the set: $\overrightarrow{EP} = [c_{angry}, c_{happy}, c_{neutral}, c_{sad}]$. We measure confidence using Support Vector Machines (SVMs). SVMs are binary maximum margin classifiers that find a separating hyperplane that maximizes the distance from the hyperplane to the points closest to the hyperplane. The outputs of SVM are class membership (± 1) and distance from hyperplane. We multiply these quantities to arrive at an approximate measure of confidence. The SVMs are trained using leave-one-speaker-out cross-validation (Figure 2).

3.2. Emotogram

The emotogram of an utterance is the set of EPs extracted over windowed regions of an utterance (See Figure 2). They provide a dynamic description of the estimated presence or absence of each emotional cue [6, 13]. This can be seen as estimating the manner in which emotion cues flow in an utterance. Emotograms are four-dimensional time-series of estimated emotion dynamics: $\overrightarrow{Emotogram} = [\overrightarrow{EP_1}, \overrightarrow{EP_2}, ..., \overrightarrow{EP_T}]$, where T represents the number of sliding windows in an utterance. In this paper, we use window lengths of 0.25, 0.5, 1.0, 1.5, and 2.0 seconds to evaluate the effect of window size on classification performance [6]. We investigated denoising techniques to mitigate subtle estimation noise. However, both Median Filtering and Kalman Smoothing techniques did not result in performance increases as compared to the *raw* emotograms. Therefore, the emotograms were not smoothed. We hypothesize that this may be because our DTW based method captures high-level emotion flow patterns, rather than the small estimation fluctuations, which would be sensitive to noise.

4. METHODS

4.1. Similarity Measurement Between Emotograms Using DTW

Our hypothesis is that the utterance-level patterns of emotion flow are informative with respect to emotion class. To test this hypothesis we measure the time-series similarity between each emotogram, our estimates of emotion flow, using DTW. DTW is a widely used technique that finds the best alignment between two time series by identifying the warping path between the two sequences that mini-

mize the difference between the sequences. DTW has been widely used in many domains including speech recognition [16] and handwriting recognition [17]. DTW captures utterance-level dynamics, rather than probabilistic transitions in frame-by-frame characteristics, which are seen in HMM modeling. DTW provides flexibility in the analysis of utterances of different lengths since it aligns time series data. In emotion data, contrasted with speech phoneme modeling, the affective data are often of highly varied length. HMMs do not offer this same flexibility because of their innate restriction to an n-state model independent of utterance length. Further, it is difficult to interpret the resulting models generated by HMMs. We present an alternative dynamic modeling technique that facilitates visualizations of affective flow, providing clear measures of emotional similarity. We propose that DTW can be an alternative strategy for emotion recognition.

We align two utterances in the emotion space defined by the emotograms using Multi-Dimensional Dynamic Time Warping (MD-DTW), presented in [18]. MD-DTW uses all emotogram dimensions to identify the best alignment between two utterances in the emotion space. We define the emotion space as $\Phi^{I \times J}$ for a descriptor of length I and dimension J, where J is the number of emotogram dimensions ($J = 4$). Let $T \in \Phi^{M \times J}$ and $L \in \Phi^{N \times J}$ be two emotograms in this space. MD-DTW computes the optimal alignment between T and L using dynamic programming ($O(MN)$) [19]. We find the optimal alignment by computing distance between the utterances. The distance measure between any two points in the series is defined as $d : \Phi \times \Phi \rightarrow \mathbb{R} \geq 0$, which can be any p-norm. We use 2-norm, the summation of the squared differences across all dimensions.

The MD-DTW algorithm populates the M by N distance matrix D according to the following equation:

$$D(i,j) = \sum_{k=1}^{K} (T(i,k) - L(j,k))^2, \quad (1)$$

where i and j represent the specific short-time estimate of the emotograms, T and L. The distance matrices can be visualized to understand the structural similarities across emotion class (Figure 3). We implemented four-dimensional DTW by modifying the one-dimensional code of [20].

4.2. *k*-Nearest Neighbor Classification Using MD-DTW

We use the k-Nearest Neighbor (k-NN) classifier to assign a final emotion class label based on the MD-DTW measure. k-NN assigns a label to a given test utterance based on the labels of its k nearest neighbors. The assigned label is a majority vote over the neighbors' labels. We select k using a 10-fold cross-validation hyper-parameter search over values 1, 3, 5, 7, 10, 30, and 50. We did this search over the combined data, which provided access to both the prototypical and non-prototypical examples and found $k = 50$.

We refer to the total framework as DTW-kNN. The algorithm is as follows. During training we calculate the DTW similarity between every pair of testing and training utterances. During testing we find the k closest neighbors to each test utterance using this DTW distance. We label the test utterance with the majority voted label of its k nearest training utterances. In both the DTW-kNN and baseline models we calculate accuracy using leave-one-speaker-out cross-validation. The final reported accuracy measures are the average of the accuracies over all 10 folds.

Table 1. *Unweighted classification accuracy (%) across different window lengths for each expression type: (A) Prototypical, (B) Non-prototypical, and (C) Combined.*

	Model	Window size (seconds)				
		0.25	0.5	1	1.5	2
A	Emotogram	66.90	66.82	67.15	68.50	67.76
	Feature	68.59	66.46	66.51	67.38	67.49
	Static EP			67.34		
B	Emotogram	53.96	55.12	55.31	55.49	55.79
	Feature	49.30	50.12	51.64	52.64	51.91
	Static EP			54.11		
C	Emotogram	63.95	64.01	64.40	64.38	64.27
	Feature	62.72	63.91	63.91	64.02	63.44
	Static EP			61.20		

4.3. Baseline Models

We evaluate DTW-kNN by comparing it to three baseline models. The first baseline model tracks emotion similarity using trajectories composed of the compressed feature space ('feature trajectories'), rather than the estimates of affective flow. We reduce our original 180 features using Principal Component Analysis (PCA). The feature dimensionality is selected using leave-one-subject-out cross-validation over compressions that reduce the features space to 4, 10, 20, 30, and 40 dimensions. The best performing model uses ten PCA features. We compare the performances of this compressed feature space to that of the affective estimates to identify the method that best allows us to capture the structure underlying emotional speech. As in the emotion flow model, we calculate the DTW similarity over each utterance, as represented by the feature trajectories, and then identify the emotion state using k-NN with k=50 (selected using hyperparameter search).

The second baseline uses static EP modeling. Static EPs are calculated in the same manner as short-time EPs. However, here the emotion is detected using utterance-level statistics (as compared to windowed statistics, e.g., over 0.25 seconds). This baseline assesses whether the dynamics contribute to our understanding of emotion class. We classify the final label of the static EP estimate using k-NN over the four dimensions ($k = 50$). In the static baseline, the k-NN classifier uses the Euclidean distance between the four-dimensional EP values of the training and test utterances.

The final baseline is a published result that modeled the dynamics of the emotograms using HMMs. HMMs fit these dynamics to an n-state model, where here $n = 3$ (with left-to-right topology) [13]. This baseline is a comparison to a result on a subset of the utterances considered in this paper (2,903 utterances vs. 3,018 utterances).

5. RESULTS

All results are reported using unweighted accuracy, the average of per class recall. This measure mitigates class imbalance in accuracy reporting. The DTW-kNN method achieves the highest performance gain for the non-prototypical utterances, the subtle utterances with only majority ground truths. The maximal accuracy of our proposed method for the non-prototypical data is 55.79% with a window size of 2 seconds. This is 3.88% higher than the feature trajectory model with the same window size, and 3.15% for the maximally accurate feature trajectory (window size 1.5 seconds). It is 1.68% higher than the accuracy of the static EP. In the prototypical data experiment, the DTW-kNN method achieves the highest accuracy of 68.50% with a

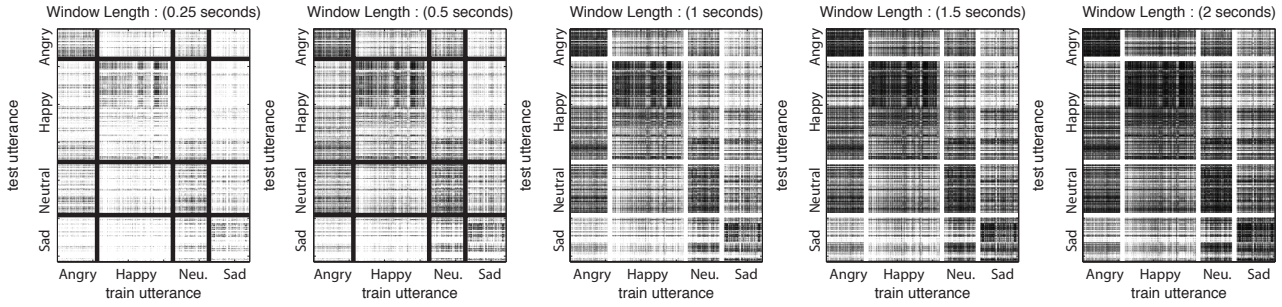

Fig. 3. *MD-DTW distance matrices between Angry, Happy, Neutral, Sad utterances (combined data). Dark represents similar patterns.*

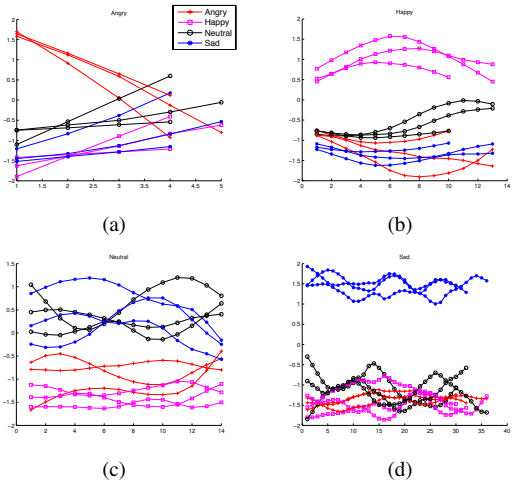

Fig. 4. A subset of utterances described by emotograms and chosen for visualization purposes, which are correctly classified by our framework: (a) Angry, (b) Happy, (c) Neutral, and (d) Sad utterances

window size of 1.5 seconds. This outperforms the feature trajectory model by 1.13% on the same window size. The maximal accuracy for the feature trajectory model is 0.09% higher than our proposed model (window size 0.25 seconds). The combined data has a maximal accuracy of 64.40% with the DTW-kNN method (window size of 1 second). This is 0.48% higher than the feature trajectory with the same window size and 0.38% higher than the feature trajectory with its maximal accuracy with window size of 1.5 seconds. It is 3.20% higher than the static EP estimate. The results are summarized in Table 1. The HMM baseline was calculated only over a window size of 0.25 seconds with an accuracy of 64.67%. This is a similar result to our proposed DTW-kNN method, 63.95%, for a window size of 0.25 seconds and suggests this restricted n-state structure may not be necessary for dynamic modeling of emotion.

6. DISCUSSION

Our results include three important findings. 1) The new dynamic modeling technique using flow pattern modeling can effectively capture the emotion dynamics. These dynamics can be used to effectively classify utterances. 2) In this framework, the secondary emotogram features outperform the compressed raw feature fluctuations, particularly in the case of non-prototypical data. This suggests that the secondary features capturing emotion flow offer a targeted compression of the emotion space. 3) For emotionally subtle utterances, our approach outperforms the baseline models.

One benefit of our method is that it provides insight into the nature of inter-class similarity. We visualize the time-series similarity distance matrix in Figure 3. The DTW distances correspond to the five window sizes of: 0.25, 0.5, 1, 1.5, and 2 seconds (left to right). The diagonal blocks of each distance matrix represent the distance between the utterances with the same emotion class. Darker regions indicate stronger similarity between the dynamics of the utterances. The dark regions on the off-diagonals of the distance matrices demonstrate that there exists confusion between *Neutral* and *Sad*, and between *Neutral* and *Angry*. This confusion mirrors the common classification error between the classes of neutrality and sadness.

The distance matrix also permits an analysis of the structural patterns that are similar. We present utterances that are similar using the MD-DTW formulation. This provides an interpretable description of typical flow patterns for each emotion class (Figure 4). In the figures, all utterances shown are correctly classified using the proposed DTW-kNN framework. The angry utterances demonstrate an interesting trend from high confidence in the presence of anger to a more mixed appraisal of emotional message. The happy trends show a peaked happiness behavior. The sad utterances display slight fluctuations in expression. The neutral utterances depicted have irregular flow patterns even though they are correctly classified (See Figure 4(c)). This can explain the relatively low classification accuracy of neutral utterances compared to that of the other emotion classes. It also supports our hypothesis that the emotion flow similarity may correspond to human emotion perception, since a label of neutrality may be provided when there exists no dominant emotional cues in an utterance.

7. CONCLUSIONS

In this paper we propose a new framework to characterize utterances based on interpretable measures of affective dynamics. We use DTW to align our affective estimates of emotion flow and then classify using the resulting distance matrix using kNN. This allows us to evaluate the discriminative power of the framework. The speaker independent experimental results are presented across five different window sizes, 0.25, 0.5, 1, 1.5, and 2 seconds for prototypical, non-prototypical, and combined data. Our results show that the proposed method outperforms the feature trajectory and the static EP baseline models. The highest improvement in our model comes from the classification of non-prototypical, or emotionally subtle, utterances. The novelty of our work is in its explicit modeling of the temporal flow patterns of emotion estimates. By taking into account the long-range dynamics of human emotion, we can have more natural and interpretable modeling techniques for emotion dynamics.

8. REFERENCES

[1] R.B. Cattell, "The structuring of change by p-technique and incremental r-technique," *Problems in measuring change*, pp. 167–198, 1963.

[2] M.W. Browne and J.R. Nesselroade, "Representing psychological processes with dynamic factor models: Some promising uses and extensions of arma time series models," *Psychometrics: A festschrift to Roderick P. McDonald*, pp. 415–452, 2005.

[3] H. Song and E. Ferrer, "State-space modeling of dynamic psychological processes via the kalman smoother algorithm: Rationale, finite sample properties, and applications," *Structural Equation Modeling*, vol. 16, no. 2, pp. 338–363, 2009.

[4] J.R. Nesselroade and P.C.M. Molenaar, "Quantitative models for developmental processes," *Handbook of developmental psychology*, pp. 622–639, 2003.

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

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

[7] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and SS. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. of INTERSPEECH Conference*, 2010, pp. 2362–2365.

[8] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 2, pp. II–1.

[9] A. Graves, *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer, 2012.

[10] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 23–30.

[11] M. Wollmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream asr framework for blstm modeling of conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4860–4863.

[12] D.B. Paul, "Speech recognition using hidden markov models," *The Lincoln Laboratory Journal*, vol. 3, no. 1, pp. 41–62, 1990.

[13] E. Mower Provost and SS. Narayanan, "Simplifying emotion classification through emotion distillation," in *Proceedings of APSIPA Annual Summit and Conference*, Dec. 2012.

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

[15] Y. Lu, I. Cohen, XS. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 301–304.

[16] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 623 – 635, dec 1980.

[17] C.C. Tappert, C.Y. Suen, and T. Wakahara, "The state of the art in online handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 8, pp. 787 –808, aug 1990.

[18] G.A. Holt, M.J.T. Reinders, and E.A. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Thirteenth annual conference of the Advanced School for Computing and Imaging*, June 13-15 2007.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[20] T. Felty, "Dynamic time warping [online]," in *MATLAB Central File Exchange*, Dec. 2005.