AFFECT RECOGNITION FROM LIP ARTICULATIONS

Rizwan Sadiq, Engin Erzin

Multimedia, Vision and Graphics Lab,
Koc University, Istanbul, TURKEY
rsadiq13,eerzin@ku.edu.tr

ABSTRACT

Lips deliver visually active clues for speech articulation. Affective states define how humans articulate speech; hence, they also change articulation of lip motion. In this paper, we investigate effect of phonetic classes for affect recognition from lip articulations. The affect recognition problem is formalized in discrete activation, valence and dominance attributes. We use the symmetric KullbackLeibler divergence (KLD) to rate phonetic classes with larger discrimination across different affective states. We perform experimental evaluations using the IEMOCAP database. Our results demonstrate that lip articulations over a set of discriminative phonetic classes improves the affect recognition performance, and attains 3-class recognition rates for the activation, valence and dominance (AVD) attributes as 72.16%, 46.44% and 64.92%, respectively.

Index Terms— Affect Recognition, Emotion Recognition, KullbackLeibler Divergence, Phoneme,Lip articulations

1. INTRODUCTION

Emotional content in a conversation helps deliver the actual meaning. Affect recognition received increased attention from researchers due to its practical applications in the fields of human-machine interaction, psychology and sociology. In psychology, emotions are represented either by a set of discrete emotional models [1] or by multi-dimensional continuous attributes [2]. Discrete emotional categories include anger, disgust, fear, happiness, sadness, and surprise, whereas a vector of affect attributes can define different levels of emotions. A common continuous emotional space defines activation, valence and dominance (AVD) attributes, which describe intensity, positivity/negativity, and degree of control.

Early studies on affect recognition have commonly used the discrete emotional models; however, recently affect recognition has been carried over the continuous attributes [3–7]. Multiple cues have been used in the literature for the purpose of affect recognition. Some studies used single modality [8, 9], while some others utilized multimodal systems, which include speech, facial expressions, head and full body movements [5, 6, 10]. In an early study, emotion recognition was investigated on phoneme level speech articulations [11]. The researchers showed that there are variations across emotional states in the spectral features at the phoneme level, especially with the vowel sounds. In a later study [12], they modeled the spectral information at the phoneme level to categorize discrete emotions, using prosodic features of speech to classify a discrete set of basic emotions. They explored the significance of different classes of phonemes (vowels, glides, nasals, fricatives and stops) for the purpose of emotion recognition, and observed only categorical emotional labels and examined the broad phonetic classes. For discrete emotion recognition from images, [13] used local binary patterns(LBP) features as input to convolutional neural network(CNN) models. They reported their results on Emotion Recognition in the Wild Challenge (EmotiW 2015) and Static Facial Expression Recognition sub-challenge (SFEW). The Significance of the lower facial area for mapping speech to facial gestures, in the presence of emotional content, was discussed by [14]. They also explored this relation for different phoneme classes (vowel, nasals, glides, stops, consonants etc.), and showed that lower facial area is more active as compared to upper face, when exposed to emotional content.

Using both audio and visual features to classify affect was also discussed in [10], where the researchers used context dependent models. They presented their results for three level affect recognition from speech only, face only and a combination of both speech and face by using Hidden Markov Models (HMM). Their findings showed that valence can be better recognized by complete facial features while activation achieves better results with speech features. The maximum accuracy of classification was 61.92% for activation using audio features and 51.36% by using complete facial features. For valence they reported a classification accuracy of 49.99% using audio features and 60.98% using facial features.

The Impact of affect on articulatory precision was investigated in [15]. Analyzing the formant position of vowel sounds under different affective states, the researchers suggested that arousal and valence have a sizable influence on formant positioning. Another contribution for affect recognition on the phoneme level was performed by [16]. Using only MFCC features extracted per utterance, they classify binary levels of arousal. They claim that using only a set of 7 vowels,

which they referred to as distinctive vowels, can also attain almost equal or little less recognition rate over word and utterance levels. In this study, we investigate the effect of phonetic classes for affect recognition from lip articulations. The affect recognition problem is formalized in discrete activation, valence and dominance attributes. A set of discriminative phonemes, which improve the affect recognition performance, is identified. The reminder of this paper is organized as follows. In Section 2, we describe the proposed methodology for the affect recognition from lip articulations. we give experimental evaluations in Section 3. Finally, conclusions are discussed in Section 4.

2. METHODOLOGY

In this section, we first define the database, which includes a rich set of lip articulation data from affective interactions. Then, the lip feature representation is given. Later, we describe a scoring function to select phonemes, which are most discriminating for the affect recognition problem. Finally, we define the affect recognition framework.

2.1. Data Set

In our analysis and experimental evaluations, we used the interactive emotional dyadic motion capture database (IEMOCAP), which is designed to study expressive human interactions [17]. The IEMOCAP is an audiovisual database, which also provides motion capture data of face, head and partially hands. It comprises spontaneous conversations between pairs of professional actors in dyadic interaction sessions. The corpus has five sessions with ten actors taking part in dyadic interactions. In each session, actors play three scripts and improvise eight hypothetical scenarios to elicit rich emotional reactions. Recordings of each session are split into clips. The total number of clips is 150 with a total duration of approximately 8 hours. Motion capture of the face has 53 markers, where 8 markers are placed on the lips. For every sentence phonetic transcriptions are also available in the database. The emotional content was annotated by human annotators on the sentence level either in categorical labels (angry, happy, excited, sad, frustrated, fearful, surprised, disgusted, neutral and other) or in dimensional descriptions of valence, activation and dominance. Value 1 denotes very low activation, dominance and very negative valence and 5 denotes very high activation, dominance and very positive valence. Those labels were assigned values between 1 and 5 and were averaged across 2-3 annotators [17].

In our experiments, we quantize AVD attribute values, A, into three levels as used in [10],

$$Q(A) = \begin{cases} A_1 & \text{if } 1 \leq A \leq 2, \\ A_2 & \text{if } 2 < A < 4, \\ A_3 & \text{if } 4 \leq A, \end{cases} \quad (1)$$

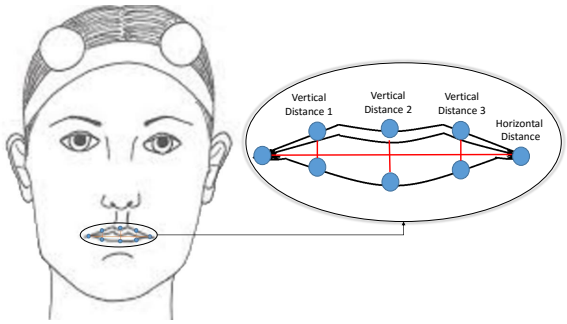

Fig. 1. Horizontal and vertical lip distances

where A_1, A_2 and A_3 are representing the low, medium and high levels of activation, dominance and valence attributes.

2.2. Feature Extraction

We computed four distance features using the x, y, z coordinates of 8 lip markers to define the frame level lip feature vector. The lip feature is comprised of one horizontal and three vertical lip distances as shown in Figure 1. We extracted statistical functionals of the lip feature vectors over temporal windows to define the segment level lip feature vectors. A total of 11 statistical functionals used were: mean, standard deviation, skewness, kurtosis, range, min, max, first quantile, third quantile, median quantile, and inter-quantile range [4]. The segment level 44-dimensional lip feature vectors, f, was computed over phoneme duration. Phoneme boundaries were used as provided by the IEMOCAP database.

2.3. Discriminative Phoneme Selection

We investigated how statistical characterization of the segment level lip features changes across different levels of affect state for each phoneme. For this purpose, we defined a symmetric KullbackLeibler divergence (KLD) of lip features f given affect state and phonetic class as

$$D_{mn}(f|p) = KLD(P(f|A_m, p), P(f|A_n, p)), \quad (2)$$

where $P(f|A_m, p)$ is the conditional probability mass function of f given affect level A_m for phoneme p. The symmetric KLD is defined over probability mass functions $X()$ and $Y()$ as

$$KLD(X, Y) = \sum_j X(j) log \frac{X(j)}{Y(j)} + \sum_j Y(j) log \frac{Y(j)}{X(j)}. \quad (3)$$

As the discrimination of affect requires significant changes of the distributions across different affect levels, we defined a cumulative distance function $S(f|p)$ for each phoneme as,

$$S(f|p) = w_p(D_{12}(f|p) + D_{23}(f|p) + D_{13}(f|p)), \quad (4)$$

where w_p is the frequency of occurrence of phoneme. Note that, the cumulative distance function $S(f|p)$ is expected to output larger distances if lip feature distribution changes largely over affective states.

2.4. Affect Classification

We used the segment level lip features to classify 3-level discrete affect values, and used the Support vector machine of LIBSVM [18] with radial basis kernel function as the classifier. Affect labels at sentence level, which are assigned by annotators, were used as ground truth to train and validate the classifiers. Classification task was performed at phoneme and sentence levels. These two approaches are briefly described in the following sub-sections.

2.4.1. Phoneme Level Classification

In the phoneme level classification, a classifier was constructed for each phoneme. The 44-dimensional segment level lip feature vectors, which were extracted over each occurrence of a phoneme, were used as input features. Over all the dataset, using the provided phoneme boundaries, all segments of each phoneme were extracted. Affect state level of each phoneme segment was extracted from the sentence level annotations. Classifiers were constructed for each phoneme and affect classification was evaluated over phoneme segments.

2.4.2. Sentence Level Classification

In the sentence level classification, we investigated two strategies: decision fusion and data fusion. In decision fusion, we applied the majority voting over the phoneme level classifier outputs. In the data fusion, we constructed segment level lip feature vectors over all duration of the sentence, and trained sentence level classifiers. On sentence level, we also investigated the use of discriminative set of phonemes to the classification performance.

3. EXPERIMENTAL EVALUATIONS

3.1. Results on Discriminative Phonemes

Over the whole database, we first computed the conditional probability mass functions for the lip features, then the cumulative distance functions were evaluated for each phoneme and for each affect attribute. We observed that more than 80% of the phonemes exhibiting higher KLD score were common in all three dimensions of affect, so we sorted the phonemes with respect to the average of distances over three affect attributes. Figure 2 and 3 respectively plots the cumulative discriminative distances of vowels and consonants according to manner and place of articulation. Note that discriminative distance gets lower as color gets darker. Hence, phonemes that

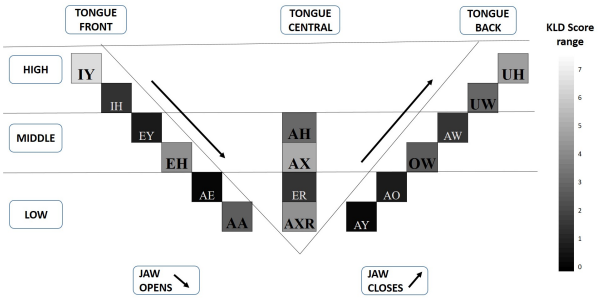

Fig. 2. Cumulative discriminative distances of vowels according to manner and placement of articulation

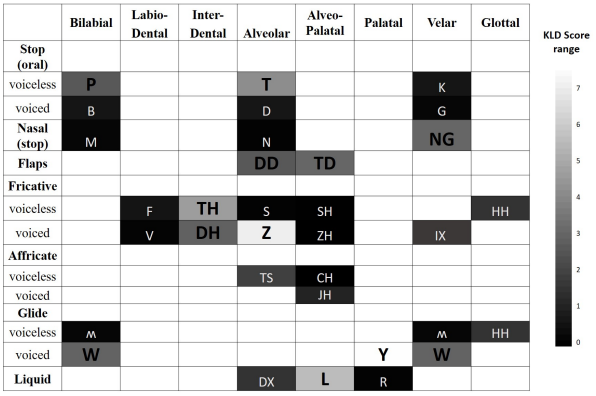

Fig. 3. Cumulative discriminative distances of consonants according to manner and placement of articulation

are discriminating affective states better with the lip features have lighter colors. Among the vowels, the least discriminative region is observed as jaw opens and tongue is at back, and we select */IY/, /EH/, /AA/, /AX/, /AXR/, /AH/, /UH/, /UW/, /OW/* as the discriminative vowels for lip driven affect recognition.

Similarly among the consonants, the voiced palatal */Y/* and the voiced alveeolar */Z/* are the top two most discriminative phonemes for the affect recognition. On the other hand, the labio-dentals */F/* and */V/* and alveo-palatals */SH/, /ZH/, /CH/, /JH/* are among the least discriminative phonemes. We select */P/, /W/, /TH/, /T/, /TD/, /DH/, /Z/, /DD/, /L/, /Y/, /NG/* as the discriminative consonants for lip driven affect recognition.

In Figure 4, we present the top 10 most discriminative and bottom 5 least discriminative phonemes for the affect recogniton task.

3.2. Classification Results

Classification experiments are organized in a leave-one-speaker-out cross validation scheme. Within our experiments, we use two performance evaluation methods: unweighted average accuracy (UA), which is the sum of all class accuracies,

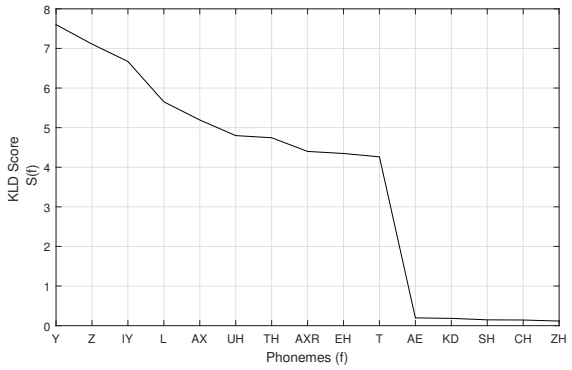

Fig. 4. KLD Score of top 10 and bottom 5 phonemes

Table 1. Comparison of sentence level weighted (WA) and unweighted (UA) classification accuracies of AVD attributes with selected phonemes and all phonemes using data fusion.

	Classification Accuracy (%)					
	All Phonemes			**Selected Phonemes**		
	A	V	D	A	V	D
WA	71.40	45.43	61.79	72.13	**46.44**	64.47
UA	40.37	38.82	38.53	41.42	**42.33**	**39.61**

divided by the number of classes, and weighted accuracy (WA), which is the number of correctly recognized labels divided by total number of occurrences. The results presented here are generated as speaker independent results. We compare classification performance of affect recognition task using all phonemes and the selected list of discriminative phonemes. Table 1 reports the sentence level classification performances using the data fusion. Similarly, Table 2 presents the sentence level classification performances using the decision fusion. Note that unweighted accuracy scores are in general lower, since number of samples from the affect state levels are unbalanced. We observe classification performance improvements with the selected list of discriminative phonemes for all affect attributes.

Table 2. Comparison of sentence level weighted (WA) and unweighted (UA) classification accuracies of AVD attributes with selected phonemes and all phonemes using decision fusion.

	Classification Accuracy (%)					
	All Phonemes			**Selected Phonemes**		
	A	V	D	A	V	D
WA	71.86	45.59	64.74	**72.16**	46.16	**64.92**
UA	39.77	35.12	38.11	**42.65**	36.33	38.98

4. CONCLUSION AND FUTURE WORK

In this paper, we investigate the role of phonemes for affect recognition under lip articulations. We observe that using only lip features can attain affect recognition performance higher than the random. Our results also suggest that a selected list of discriminative phoneme articulations can better classify all affect attributes using only the lip features. These results are encouraging for the use of lip articulations in multimodal affect recognition systems.

We also deliver an analysis of discriminative phonetic classes for the lip driven affect recognition as a function of manner and place of articulation. Extensive studies in this direction have potential to improve contribution of lip modality to the affect recognition systems.

5. REFERENCES

[1] Silvan S Tomkins, *" Affect Imagery Consciousness: Volume I: The Positive Affects "*, vol. 1, Springer publishing company, 1962.

[2] Harold Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, pp. 81, 1954.

[3] Hatice Gunes, Maja Pantic, and J Vallverdú, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, pp. 68–99, 2010.

[4] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[5] Elif Bozkurt, Shahriar Asta, Serkan Özkul, Yücel Yemez, and Engin Erzin, "Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3652–3656.

[6] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[7] Hossein Khaki and Engin Erzin, "Continuous emotion tracking using total variability space," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan, "Using neutral speech models for

emotional speech analysis," in *Interspeech*, 2007, pp. 2225–2228.

[9] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[10] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[11] Lea Leinonen, Tapio Hiltunen, Ilkka Linnankoski, and Maija-Liisa Laakso, "Expression of emotional–motivational connotations with a one-word utterance," *The Journal of the Acoustical society of America*, vol. 102, no. 3, pp. 1853–1863, 1997.

[12] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition based on phoneme classes," in *Interspeech*, 2004, pp. 205–211.

[13] Gil Levi and Tal Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 503–510.

[14] Carlos Busso and Shrikanth S Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[15] Martijn Goudbeek, Jean-Philippe Goldman, and Klaus R Scherer, "Emotion dimensions and formant position," in *Interspeech*, 2009, vol. 2009, pp. 1575–1578.

[16] Bogdan Vlasenko, Dmytro Prylipko, Ronald Böck, and Andreas Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483–500, 2014.

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

[18] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.