ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

ANALYSIS AND MODELING OF AFFECTIVE AUDIO VISUAL SPEECH BASED ON PAD EMOTION SPACE

Shen Zhang, Yingjin Xu, Jia Jia, and Lianhong Cai

Department of Computer Science and Technology, Tsinghua University

{zhangshen05, xuyj03, jiajia}@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

ABSTRACT

This paper analyzes acoustic and visual features for affective audio-visual speech based on PAD (Pleasure-Arousal-Dominance) emotion space. The selected acoustic features include F0 maximum, F0 minimum, duration and energy. A set of Partial Expression Parameters (PEP) is proposed as visual features to describe affective facial movement on talking face. This paper explores the connection between PAD emotion space and acoustic/visual features respectively. The variation of acoustic features is predicted by PAD values, and a PAD-PEP mapping function for facial expression synthesis is built. Experimental result shows that PAD could be properly applied in describing emotional state as well as predicting the acoustic/visual features for affective audio-visual speech synthesis.

Index Terms: emotion, acoustic features, facial expression, talking face, speech synthesis

1. INTRODUCTION

The audio-visual speech synthesis provides a face-to-face human computer speech interface. However, due to lack of emotion, it is still not as affective as real human speech communications. Shouting loudly usually means "I am very angry", screaming with high pitch indicates excited or scared and muttering is accompanied with discontent or sorrow. Unfortunately such kind of affective information is not conveyed in traditional audio-visual-speech synthesis where the output is combination of neutral speech signal and dull talking face. According to Mehrabin's study [1], the affective audio-visual speech interface can facilitate the delivery of nonverbal information. For emotional speech, a variety of approaches are proposed [2] and one of the open problems is modeling and modification of acoustic features. For expressive talking face, facial expression is studied to embed kinds of emotions in visual speech.

In this paper, both acoustic and visual features are analyzed quantitatively based on PAD emotion space. Human emotional state is not classified into categories but described by three-dimensional (P, A, D) parameters. An affective speech corpus and a facial expression database are established respectively, with PAD annotation obtained by subjective evaluation. A statistic model for predicting acoustic feature variation by PAD, is trained by the joint-GMM algorithm, and a polynomial emotion-expression mapping function is trained to capture the relation between PAD and Partial Expression Parameters (PEP) which are proposed to depict the local facial movement. Based on the above analyses, we aim to build an affective audio-visual speech synthesizer, where the input is (P, A, D) value, which indicates the target emotional state, and synthetic neutral speech, while the output is affective speech and expressive talking face.

The rest of this paper is as follows. Section 2 introduces the PAD emotion space. The training of GMM-based model for predicting acoustic variation by PAD is presented in section 3. In section 4 we describe the emotion-expression mapping function. Finally, we conclude our work and future direction on affective audio-visual speech synthesis.

2. PAD EMOTION SPACE

The PAD emotion space is not only a scale for describing human emotional state, but can help us to build a clear connection between high-level human perception and low-level acoustic/visual signals. The PAD emotion space is proposed by Mehrabin [3] for psychological research. According to his theory, emotions are not limited to isolated categories but can be described along three nearly independent continuous dimensions: Pleasure-Displeasure (P), Arousal-Nonarousal (A), and Dominance-Submissiveness (D). Emotional states are not described by a set of categories or descriptive words, but denoted as points in three-dimensional PAD emotion space. In this way, different emotions can be distinguished quantitatively along the P-A-D dimensions respectively.

A subjective method [4] for evaluating PAD values by a 12-item questionnaire is adopted to obtain PAD annotation for affective speech and facial expressions. As shown in Table 1, there are 12 pairs of descriptive word and for each pair of the words (Emotion-A and Emotion-B), which are just like two ends of a scale, the annotator is required to choose one of them that better describes the affective speech or facial expression with a 9 level score varying from -4 to +4. The P, A and D values are then calculated from this questionnaire using the method described in [3], and are normalized to [-1, +1]. The 12-item PAD scale has been proved as a versatile psychological measuring instrument which is capable of adapting to a variety of applications including emotion annotation.

The overview of our work on acoustic and visual features analysis and modeling is shown in Figure1. We aim to model the correlation between PAD emotion space and acoustic/visual feature vector space, and then try to predict the affective acoustic features and visual features for affective talking face, and the integration of them can produce an affective audio-visual speech synthesizer.

Table 1. *12-item questionnaire for PAD evaluation*

Emotion-A	Emotion-B	Emotion-A	Emotion-B
Angry	Activated	Cruel	Joyful
Wide-awake	Sleepy	Interested	Relaxed
Controlled	Controlling	Guided	Autonomous
Friendly	Scornful	Excited	Enraged
Calm	Excited	Relaxed	Hopeful
Dominant	Submissive	Influential	Influenced

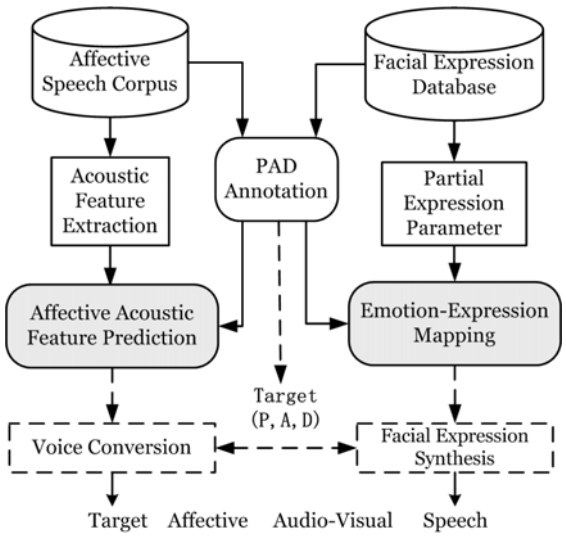


Figure 1.*Overall framework of acoustic/ visual features analysis*

3. PREDICTING MODEL OF ACOUSTIC VARIATION

3.1 Affective Speech Corpus

10 typical emotional states are selected to build the affective speech corpus, and 5 paragraphs are designed for each state. Totally 55 paragraphs (including neutral state) are used for speech recording. Four speakers (2 female and 2 male) are invited for audio recording and 338 affective speech sentences are collected. PAD value is annotated for each sentence by 12-item subjective questionnaire in Table 1. The average PAD value for each emotional state is shown in Table 2. Four time-domain acoustic features are extracted for analysis, including the F0-max, F0-min, duration and energy.

Table 2. *Average PAD for speech sentences with 10 emotions*

Emotion	P	A	D
Angry	-0.90	0.79	0.95
Disdainful	-0.46	-0.36	0.69
Disgust	-0.59	0.48	0.55
Happy	0.68	0.68	0.43
Relax	0.21	-0.59	0.24
Surprise	0.16	0.74	-0.03
Anxious	-0.36	0.74	-0.08
Docile	0.33	-0.36	-0.44
Sad	-0.42	-0.44	-0.64
Scared	-0.44	0.79	-0.73

3.2 Build Predicting Model using Boosting-GMM

3.2.1 Preliminary work

Strong connections exist among different acoustic features. Kain et.al.[5] proposed a voice conversion algorithm using the Gaussian Mixture Models with joint density estimation, which considers the correlation between input and output as well as each dimensions of input/output. For affective speech conversion, we aim to build a predicting model which considers the connections among acoustic features. The input of the predicting model is PAD parameters (i.e. emotional state), and the output is acoustic variation between neutral state and emotional state. As a preliminary work, we use joint-GMM to predict the acoustic feature for "neutral state (P=0,A=0,D=0)" since the average PAD for neutral speech in our corpus is not absolute (0, 0, 0) but (0,-0.3,-0.12). The predicted acoustic features of "neutral state" is used to calculate the acoustic variation and then the GMM-based acoustic variation predicting model is built up.

3.2.2 Boosting-GMM Algorithm

In order to improve the prediction accuracy, we adopt Ada-Boost algorithm [6] in our GMM based predicting model, which we call Boosting-GMM, to get a strong model by learning from errors. Since multiple training-sets are required for Boosting-GMM to build multiple predictors, we employ the "Bagging" algorithm [7] to derive multiple training-sets from our corpus. There are two steps in our algorithm, first to train multiple weak predictors, and then to predict emotional acoustic features by PAD parameters.

- Training multiple predictors

For N acoustic features, $N+1$ predictors are trained in Boosting-GMM, including one main predictor and N support predictors. The main predictor is the acoustic variation predicting model built in 3.2.1. The support predictors aim to reduce the prediction error of corresponding features as much as possible. Let K is the size of training set T, then:
a) Main predictor M is built on the original training set $T=\{t_1,t_2,\ldots,t_K\}$ using the joint-GMM algorithm;
b) Calculate the prediction error $E=\{e_{i1},e_{i2},\ldots,e_{iK}\}$ for the i^{th} acoustic feature (i=1,2,…,N);
c) Take repeated bootstrap sample from T with each sample t_j being selected at the probability of e_{ij} (j=1,2,..,K) (equal to the prediction error) and get the new training set $T_i=\{t_{i1},t_{i2},\ldots,t_{iK}\}$;
d) The i^{th} support predictor M_i is built on the new training set T_i using the joint-GMM algorithm.

- Predicting acoustic feature

In order to get more accurate predicting result by multiple predictors, an error estimation model M_e is built, where the input is PAD and the predicted acoustic values by $N+1$ predictors, and output is the estimation error of each predictor. For the i^{th} acoustic features P_i, the predicted values are $P_{i0},P_{i1},\ldots,P_{iN}$. The estimation error estimated by M_e is $E_{i0},E_{i1},\ldots,E_{iN}$. The predictor with minimum estimation error is thus selected: $P_i=P_{ij}$ where E_{ij}=min $\{E_{i0},E_{i1},\ldots,E_{iN}\}$. The Boosting-GMM is evaluated by its prediction accuracy as defined in Equation 1, where K is the size of training set, t_j is the acoustic feature of the j^{th} sample while p_j is the predicted value. Experimental results in Table 3 show that Boosting-GMM can gain higher prediction accuracy than joint-GMM.

$$P = 100 \times (1 - \frac{1}{K}\sum_{j=1}^{K}\left|\frac{t_j - p_j}{t_j}\right|) \quad (1)$$

Table 3. *Prediction accuracy of Boosting-GMM and joint- GMM*

Accuracy	*F0 max*	*F0min*	*Energy*	*Duration*	***AVERAGE***
Joint-GMM	88.9%	82.0%	94.8%	90.6%	**89.0%**
Boosting-GMM	90.5%	85.9%	96.4%	94.0%	**91. 7%**

3.2.3 Acoustic variation analysis based on PAD

Four predicting models are built for each speaker in our affective corpus. Since people have different ways to express emotion, even speech with the same emotional state will result in different acoustic variations. We attempt to find some general laws on acoustic variation with the PAD as a bridge to connect the acoustic feature and emotional state. Four acoustic features including F0-max, F0-min, duration and energy are predicted for each emotional state. Table 4

illustrates the statistics of the variation of these acoustic features as well as the F0-range which can be calculated by the F0-max and F0-min.

According to the statistics, it is found that the *Arousal* dimension of PAD emotion space is close related with the variation of acoustic features. For emotional state with positive A value, the F0 get higher, the energy become larger, while the duration is shorter, and vice versa. This phenomenon shows that the *Arousal* dimension is positive related to the variation of F0, and the same conclusion is obtained in Pereira's [8] study.

Table 4. *Variation of acoustic features for 10 emotion speech*

Emotion	*F0 max*	*F0 min*	*Energy*	*Duration*	*F0Range*
Angry	40±13%	33±17%	19±2%	-23±11%	6±10%
Disdainful	-3±7%	-15±20%	-4±6%	5±11%	20±36%
Disgust	19±21%	-1±34%	10±3%	-8±7%	29±37%
Happy	28±15%	19±13%	14±8%	-8±3%	8±12%
Relax	-24±6%	-23±14%	-12±2%	12±11%	2±18%
Surprise	22±13%	27±22%	12±5%	-16±4%	-3±15%
Anxious	26±15%	31±27%	16±5%	-23±7%	-2±13%
Docile	-21±8%	-18±22%	-12±3%	5±6%	1±19%
Sad	-20±13%	-1±18%	-6±10%	11±14%	-17±22%
Scared	26±12%	33±25%	15±9%	-17±19%	-3±18%

4. EMOTION-EXPRESSION MAPPING

4.1 Partial Expression Parameters

For visual speech synthesis, the Facial Animation Parameters (FAPs) [9] is commonly used for geometric facial deformation. The FAPs can also be used to generate various facial expressions. However it is very complicated for facial expression synthesizer to manipulate the FAPs directly since only the motions (e.g. translations and rotations) of single facial feature points are defined in FAPs.

Table 5. *PEP definition by local facial movement*

Face Region	PEP code	Description [0,-1]/[0,1]
Eye-brow	1.1(L/R)	Eyebrow lower down/raise up
	1.2(L/R)	Eyebrow Relax/Squeeze
	1.3(L/R)	In the shape of "\ /" or "/ \"
Eye	2.1(L/R)	Eye-lid Close/Open
	2.2(L/R)	(Eyeball) look right/left
	2.3(L/R)	(Eyeball) look up/down
Mouth	3.1	Close/Open mouth
	3.2	Mouth-corner bent down/up
	3.3	Mouth sipped/protruded (pout)
	3.4	Mouth stretched in/ out
Jaw	4.1	Jaw move up/lower down
	4.2	Jaw move right/left

Based on our previous work on FAP-driven facial expression synthesis [10], we propose the Partial Expression Parameters (PEPs) to depict the expression movements within local face region, such as mouth-bent, eye-open and eyebrow-raise etc. Detailed description for PEP can be found in Table 5. The value of the PEP ranges in [-1, 1] corresponding to the continuous facial expression movement. For better understanding, the partial expression movement of *mouth-bent* (PEP 3.2) is illustrated in Figure 2. The PEP is adopted as visual feature in our analysis on facial expression.

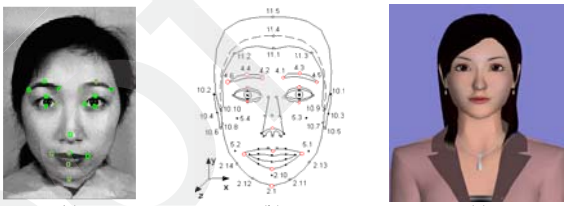
(a) PEP3.2 ranges from 0 to 1

(b) PEP3.2 ranges from 0 to -1

Figure2. *Mouth-bent with PEP 3.2 ranging in [-1, 1]*

4.2 Pseudo Facial Expression Database

A pseudo facial expression database is created. Here, "*pseudo*" means the database is not real human expression but the cartoon-like expression on a talking face. We choose the Japanese Female Facial Expression (JAFFE) database [11] as reference to build our pseudo database. The JAFFE database contains 213 images with 10 Japanese females posing three or four examples for each of seven expressions: *Neutral*, *Happy*, *Sad*, *Surprise*, *Angry*, *Disgust* and *Fear*. For each image we annotated 18 facial points manually according to MPEG-4 facial definition points (FDPs) [9] as shown in Figure 3(a) and 3(b). 12 PEP parameters are extracted by measuring the movement of above facial points as described in [12]. Pseudo facial expression is then synthesized on talking face as shown in Figure 3(c).



(a) (b) (c)

Figure 3. *Annotated facial points (a), FDPs (b), Talking face (c)*

4.3 Emotion-Expression Mapping

4.3.1 PAD annotation for pseudo expression

For each synthetic expression image in the pseudo expression database, we annotate the *P*, *A* and *D* value using the 12-item questionnaire.

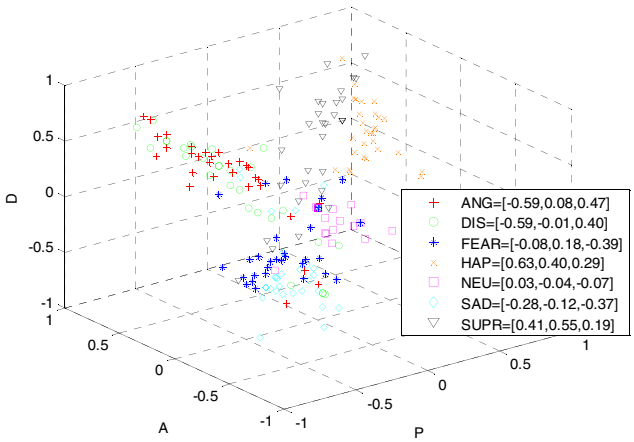


Figure 4. *PAD distribution of pseudo expression (with average PAD value denoted on the right)*

The distribution of PAD annotation is shown in Figure 4. We can see that the PAD annotations for each emotional state are distributed in nearly different areas with the exception of *Angry* and *Disgust*. This may be understandable because that some commonalities of facial movements are shared for *Angry* and *Disgust* as reported in [10].

4.3.2 PAD-PEP Mapping

Our work focus on exploring the relationship between high-level emotion description and mid-level expression configuration, in other words, we attempt to find a proper emotion-expression mapping function between PAD and PEP, we have tried the polynomial function with the first and second order as well as non-linear exponential function. As the experimental result reveals, it is the second order polynomial function that receive the best fitting result as shown in equation 2. Where ***PEP*** is the PEP vector of expression configuration, and ***E*** is the PAD vector [*P, A, D*], $\boldsymbol{E}^2$ is also a vector with each element the square value of its counterpart in ***E*** respectively, i.e. [P^2, A^2, D^2], $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding coefficient matrix, $\boldsymbol{\delta}$ is the constant offset vector.

$$\boldsymbol{PEP} = \alpha \bullet \boldsymbol{E}^2 + \beta \bullet \boldsymbol{E} + \delta \quad (2)$$

In order to reduce the limitation caused by the small database, the K-fold cross-validation method (K=10) is employed to train the PAD-PEP mapping function. By such division scheme, we are able to capture the common facial expression movement shared by different people as much as possible. The coefficients of the PAD-PEP mapping function is estimated using the least square errors method. It should be noticed that each dimension of PEP (e.g. PEP_i) is estimated separately with the same form as shown in equation 2. In the K-fold cross-validation training process (k=10), there are 10 iterations corresponding to 10 validating subset. In each iterations, we calculate the correlation coefficients of real and estimated data as criteria to evaluate the fitting performance of the trained function. The trained function with the average fitting performance among all the 10 iterations is chosen as the final mapping function.

4.3.3 Synthetic result and perceptual evaluation

The pseudo expressions for 6 basic emotional state (*happy*, *surprise*, *sad*, *scared*, *angry*, and *disgust*) is synthesized by the PEP predicted from the PAD-PEP mapping function. The input PAD values for the six basic emotional state are randomly selected from the annotation of pseudo database. 14 subjects are invited in a perceptual test to evaluate the synthetic expression with the 12-item PAD questionnaire. The input PAD values and the PAD perception values for synthetic expression are summarized in Table 6.

Table 6. *Result of PAD evaluation on synthetic expression.*

Expression	Input PAD			Evaluated PAD		
	P	A	D	P	A	D
Happy	0.55	0.24	0.28	0.42	0.12	0.10
Surprise	0.34	0.34	0.04	0.36	0.45	-0.05
Sad	-0.18	0.03	-0.14	-0.01	-0.26	-0.27
Scared	-0.19	0.26	-0.13	0.01	-0.04	-0.25
Angry	-0.40	0.22	0.12	-0.17	0.02	-0.08
Disgust	-0.36	0.08	0.13	-0.56	0.15	0.44

The correlation coefficients between input PAD and evaluated PAD are 0.89(P), 0.68(A) and 0.70(D) respectively. The experimental results shows that the proposed PAD-PEP mapping function is efficient in facial expression synthesis for talking face.

5. CONCLUSION AND FUTURE WORK

In this paper we analyze acoustic and visual features for affective audio-visual speech based on PAD emotion space. The analysis explores the connection between PAD emotion space and affective speech and facial expressions respectively. A statistic model for predicting acoustic features variation by PAD is built using the boosting-GMM algorithm, and a PAD-PEP mapping function is built to synthesize facial expression on 3D talking avatar. Experimental result shows that the PAD could be properly applied in describing emotional state and predicting the acoustic and visual features for affective audio-visual speech synthesis.

Our future work aims to integrate the acoustic predicting model in the voice conversion system which converts the neutral synthetic speech into affective speech, and to implement the facial expression synthesizer for producing affective talking face. Based on these work, we can build an affective text-to-audio-visual-speech system.

6. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (60433030,60418012) and the National Basic Research Program of China (973 Program) (No. 2006CB303101).

7. REFERENCES

[1] Mehrabian, A., "Nonverbal communication", Aldine-Atherton, Chicago, Illinois, 1972.

[2] Schröder, M., "Emotional Speech Synthesis:A Review", Proc. Eurospeech 2001, Aalborg, Vol. 1, pp. 561-564, 2001.

[3] Mehrabian, A., "Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament", Current Psychology: Developmental, Learning, Personality, Social, Volume 14, 261-292, 1996.

[4] Li, X.M., Zhou, H.T., Song, S.Z., Ran, T., Fu, X.L., "The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales", Int. Conf. on Affective Computing and Intelligent Interaction (ACII), 513-518, 2005.

[5] Kain, A., Macon M. W., "Spectral voice conversions of text-to-speech synthesis", Proc. of ICASSP'98, 1: 285-288, 1998.

[6] Freund,Y., Schapire ,R. E., "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, 55(1):119–139,August 1997.

[7] Breiman, L., "Bagging Predictors", Machine Learning, Volume 24, Number 2, August, 1996

[8] Pereira, C., "Dimensions of emotional meaning in speech", Presented at ISCA Workshop on Speech and Emotion. 2000.

[9] Motion Pictures Expert Group, ISO/IEC 14496-2: 1999/Amd. 1: 2000(E). International Standard, Information Technology – Coding of Audio-Visual Objects. Part 2: Visual; Amendment 1: Visual Extensions.

[10] Wu, Z.Y., Zhang, S., Cai, L.H., Meng, H.M., "Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a Three-dimensional Avatar", Proc. Int. Conf. on Spoken Language Processing (ICSLP) 1802-1805, 2006.

[11] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J., "Coding. facial expressions with gabor wavelets", Proc. of the 3rd IEEE Conf. on Face and Gesture Recognition, 200–205,1998

[12] Zhang, S., Wu, Z. Y., Meng, H. M., Cai, L. H., "Facial Expression Synthesis using PAD Emotional Parameters for a Chinese Expressive Avatar", Int. Conf. on Affective Computing and Intelligent Interaction, LNCS 4738, pp. 24–35, 2007.9.