Suppressed Negative-Emotion-Detecting Method by using Transitions in Facial Expressions and Acoustic Features

Joji Uemura, Kazuya Mera, Yoshiaki Kurosawa, and Toshiyuki Takezawa

Graduate School of Information Sciences, Hiroshima City University
3-4-1, Ozuka-Higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
uemura@ls.info.hiroshima-cu.ac.jp, {mera, kurosawa, takezawa}@hiroshima-cu.ac.jp

Abstract

We propose a method of detecting suppressed/concealed negative emotion during compliment utterance. When people suppress/conceal emotions, very brief facial expressions called "micro expression" often appear. In order to detect such short-duration facial expression, we propose 90 features calculated from the contours of likelihood ratios for each of the five emotions (happiness, sadness, surprise, anger, and neutral). Likelihood ratios are calculated from still images in a video every 100 milliseconds. Furthermore, 384 acoustic features are calculated for multimodal analysis. Three machine learning classifiers by Support Vector Machines were constructed by using feature sets consist of facial-expression-transition, voice, and both of them, and the classifiers were evaluated how they can detect insincere compliments in Japanese. The experimental results indicate that the feature set including both of facial-expression-transition and voice was the most superior. Its precision and recall of insincerity detection and the total accuracy rate were 0.50, 0.44, and 0.64, respectively. The results were better than the annotation results by non-expert participants.

Keywords: emotion recognition, transition of facial expression, micro expression, acoustic feature, suppressed emotion

1. Introduction

There are several human-computer interaction systems considering emotions (Mera et al., 2010; DeVault et al., 2014; Li et al., 2017), and some social robots, such as Pepper, have functions to recognize and express emotions. Considering that a user's emotion is one of the most important factors for comfortable communication, Takahashi et al. constructed a dialogue system to make reaction considering a user's emotion and evaluated users' impressions of the system (Takahashi et al., 2017). The experimental results indicated that users felt that the system was more likable and enjoyable than dialogue system that do not have a function to estimate user's emotion.

However, people do not always express their emotions because of modesty, embarrassment, suppression of aggressive emotion, evasion of responsibility, and so on. However, if a human-computer interaction system can detect leakage of suppressed/concealed emotion, it can provide more sensible and sympathetic reception. For example, when a person gloomily says "I'm fine," people will respond not by saying "That's wonderful" but "Really?" or "Are you OK?".

There are various approaches to detection of deceit. (Hancock et al., 2008) investigated the liar's and the conversational partner's linguistic style across truthful and deceptive dyadic communication in a synchronous text-based setting. (Hirschberg et al. 2005) proposed a method to distinguish deceptive from non-deceptive speech on acoustic/prosodic and lexical features. (Zhang et al. 2007) proposed an automatic deceit detection method from involuntary facial expressions. (Rigoulot et al., 2014) investigated what brain processes allow listeners to detect when a spoken compliment is meant to be sincere or not.

However, (Ekman, 2009) argues that "*There is no sign of deceit itself ... There are only clues that the person is poorly prepared and clues of emotions that don't fit the person's line*. *Line* indicates "words" in this quotation. He also argues that words, voice, body, and facial expression include the clues of emotion. Therefore, we assume that assembling multimodal information is more effective than single-modal approach to detect suppressed emotions.

In this paper, we propose a method of detecting suppressed negative emotion from words, voice, and facial expressions by using a machine-learning classifier. To detect insincere compliments, the method first detects verbal positive-emotion expressions from uttered sentences. Then, it detects the clues of insincerity from the transitions in facial expressions and acoustic features of a voice. Finally, it determines whether the speaker said the compliment insincerely by using the machine-learning classifier.

2. Suppressed Negative-Emotion-Detecting Method from Voice and Facial Expression

The proposed method detects suppressed negative emotion from words, voice, and facial expressions by using a machine-learning classifier. The method attempts to detect clues of an emotion that do not fit the content of the utterance. Usually, a speaker arouses positive emotion when he/she admires someone or something. Therefore, the proposed method estimates whether the speaker's compliment is insincere when it detected his/her negative emotion that does not fit the speaker's words.

2.1. Method Overview

Figure 1 shows an overview of our proposed method. When a person utters, the utterance is first translated into text through a speech-recognition process. The method then identifies whether the utterance is a compliment by using the "Compliment-Expression Database." When the text of the utterance matches an entry in the Compliment-Expression Database, the utterance is determined as "a compliment." The matching process considers not only

exact matches but also partial matches. If the utterance is determined as "not a compliment," the method no longer works. If the method determines the utterance as a "compliment," it starts to evaluate whether the compliment is insincere.

To evaluate a compliment utterance, the method attempts to find clues of negative emotion from the transitions in facial expressions and acoustic features of the voice. To detect micro and long-duration expressions and recognize superimposed expressions, the features relating to the transitions in facial expressions are calculated from not only all of the voice but also a part of the voice. The acoustic features mainly relate to pitch and power. The details of the feature-calculation method from facial expressions and voice are given in Sections 2.3 and 2.4, respectively. Finally, the machine-learning classifier determines whether the uttered compliment is insincere. A support vector machine (SVM) is used as the classifier.

2.2. Compliment-Expression Database

We created the Compliment-Expression Database to detect compliment utterances. To create the database, compliment expressions for various situations were first collected from a website created by a psychological counselor (Lifedata, 2015). Some of the collected expressions were "思いやりがあるね *(You are thoughtful.)*", "センスがいいね *(You have a good fashion sense.)*", and "素晴らしい *(Amazing!)*". Table 1 lists the numbers of collected compliment expressions for each topic.

To detect variations in spoken expressions, the following three types of expression forms were added to the database:

- Those with auxiliary verbs removed
- Those with only nouns and adjectives
- Those with only adjectives.

By applying these arrangements to the collected expressions, some entries did not make sense because of too much reduction. These unnecessary entries were removed manually, and overlapped entries were removed automatically. Finally, 754 expressions were registered to the Compliment-Expression Database.

2.3. Facial-Expression-Transition Features

Ekman introduces many clues to detect concealed or falsified emotion. Tables 2 and 3 show the facial clues in which an emotion is concealed and falsified with reference to Ekman's study (Ekman, 2009), respectively. **Manipulators** include all those movements in which one body part grooms, massages, rubs, holds, pinches, picks, scratches, or otherwise manipulates another body part. The actions can be also performed with the tongue against cheeks, teeth slightly biting lips, or leg against leg. Manipulators may be of very short duration or they may go on for several minutes. **Micro expressions** provide a full picture of the concealed emotion, but so quickly that it is usually missed. A micro expression flashes on and off the face in less than one-quarter of a second. **Squelched expressions** are much more common than micro expressions. As an expression emerges, the person seems to become aware of what is beginning to show and interrupts the expression, sometimes also covering it with

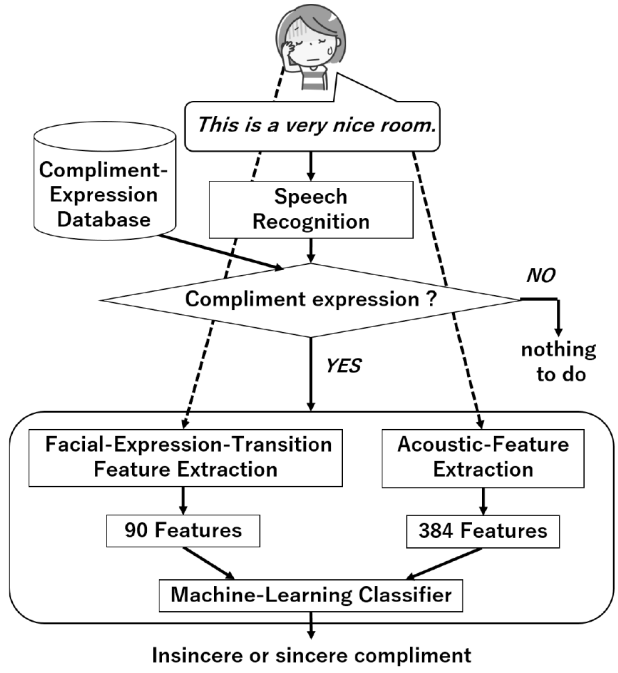

Fig. 1: Overview of proposed method

Situation	Complimenting Topic	Number of Expressions
to male	Appearance	69
	Mood	12
	Character	142
	Action	44
	Ability	57
to female	As a whole	18
	Appearance	118
	Mood	44
	Personality	73
	Action	16
	Ability	57
to child	As a whole	8
	Appearance	24
	Mood	19
	Personality	61
	Action	49
	Ability	64
business scenario	As a whole	10
	Mood	6
	Personality	42
	Action	57
	Ability	90
Total		1,080
Total (removed overlapped entries)		754

Table 1: Numbers of compliment expressions

another expression. The squelched expression is interrupted, and the expression is not always fully displayed, but it lasts longer than a micro expressions, and the interruption may be noticeable. **Reliable muscles,** such as of the forehead and eye, cannot be used to make in false expressions, and it is difficult to control them to conceal emotion expression.

To consider these facial clues, we define features particularly related to the duration of emotion expression and superimposed multiple facial expressions. Therefore,

the process to extract facial-expression-transition feature in Fig. 1 is used to analyze still facial images every 100 milliseconds. Our method uses OKAOVision SDK (OMRON, 2014) to estimate expressed emotion from a facial image. The analysis results are outputted as likelihood ratios for each of the five emotions (happiness, sadness, surprise, anger, and neutral). Therefore, five contours indicating the transitions in likelihood ratios of the five emotions are obtained from a video. The following six features are then calculated for each contour:

- **max**: The maximum value of the contour
- **min**: The minimum value of the contour
- **range** = max - min
- **maxPos (absolute)**: The absolute position of the maximum value (in frames)
- **maxPos (relative)**: The relative position of the maximum value (in frames)
- **amean**: The arithmetic mean of the contour.

Each emotion contour is then equally divided into three periods to detect short-duration facial expression. Then, four features (max, min, range, and amean) are calculated from the partial contours. A total of 90 features of transitions in facial expressions are calculated from a video.

2.4. Acoustic Features of Voice

To estimate emotions exactly, using multi-modal information is effective. Therefore, the proposed method also uses acoustic features of the video in addition to 90 features, as discussed in Section 2.3.
Table 4 lists the acoustic clues of concealed expression with reference Ekman's study (Ekman, 2009). Ekman did not introduce any acoustic clues of falsified emotion. Note that pause, pitch, power, and speech rate are important for detecting concealed emotion.
Therefore, our method uses the **IS09 feature set** extracted using openSMILE (Eyben et al., 2010) as acoustic features for our machine-learning classifier. This feature set consists of 384 static-feature values and refers to the following 16 low-level descriptors (contours):

- **pcm_RMSenergy**: Root-mean-square frame energy
- **mfcc**: Mel-Frequency cepstral coefficients 1-12
- **pcm_zcr**: Zero-crossing rate of time signal (frame-based)
- **voiceProb**: The voicing probability computed from the Autocorrelation Function
- **F0**: The fundamental frequency computed from the cepstrum.

Because a 1st order delta coefficient (differential) of each contour is calculated for each contour, the number of contours becomes 32. The following 12 features are calculated for each contour:

- **max**: The maximum value of the contour
- **min**: The minimum value of the contour
- **range** = max – min
- **maxPos**: The absolute position of the maximum value (in frames)

Clue of Concealment	Information Retrieved
Manipulators increase	Negative Emotion
Micro expressions	Any specific emotions
Squelched expressions	Specific emotions; or may only show that emotion was interrupted but not which one
Reliable facial muscles	Fear or sadness
Increased blinking	Emotion, not specific
Pupil dilation	Emotion, not specific
Tears	Sadness, distress, uncontrolled laughter
Facial reddening	Embarrassment, shame, or anger; maybe guilt
Facial blanching	Fear or anger

Table 2: Facial clues of concealed emotion

False Emotion	Information Retrieved
Fear	Absence of reliable forehead expression
Sadness	Absence of reliable forehead expression
Happiness	Eye muscles not involved
Negative emotions	Increased manipulators
Any emotion	Asymmetrical expression, onset too abrupt, offset too abrupt or jagged

Table 3: Facial clues of falsified emotion

Clue of Concealment	Information Retrieved
Pauses and speech errors	Verbal line not prepared; or, negative emotions, most likely fear
Voice pitch raised	Negative emotion, probably anger and/or fear
Voice pitch lowered	Negative emotion, probably sadness
Louder, faster speech	Probably anger, fear, and/or excitement
Slower, softer speech	Probably, sadness and/or boredom

Table 4: Acoustic clues of concealed emotion

- **minPos**: The absolute position of the minimum value (in frames)
- **amean**: The arithmetic mean of the contour
- **linregc1**: The slope (m) of a linear approximation of the contour
- **linregc2**: The offset (t) of a linear approximation of the contour
- **linregerrQ**: The quadratic error computed as the difference in the linear approximation and actual contour
- **stddev**: The standard deviation of the values in the contour
- **skewness**: The skewness (3rd order moment)
- **kurtosis**: The kurtosis (4th order moment).

A total of 384 features are calculated from a voice. The acoustic features are extracted for entire recording because estimating emotions from small fragments of a voice is much more difficult than that from a static facial image.

3. Experiment

We conducted an experiment to evaluate how the proposed method detects insincere compliments.
openSMILE (Eyben et al., 2010) was used for acoustic feature extraction, facial-expression-transition features were calculated using OKAOVision (OMRON, 2014) and the SVM in statistical computer environment R

(Ihaka and Gentleman, 1996) was used as the machine-learning classifier. Three types of feature sets (facial-expression-transition features only, acoustic features only, both facial-expression-transition features and acoustic features) were compared.
Section 3.1 explains how the experimental data were prepared, and Section 3.2 describes the experimental results.

3.1. Sincere/Insincere Compliment Examples

The proposed method was evaluated using videos of a person complimenting an object sincerely and insincerely. To record sincere/insincere compliment scenes as video, seven participants (undergraduate or graduate university students aged 21 to 24) were asked to praise objects shown in photographs **even if they were not impressed.** All of the participants were males to control the degree of emotion expression because women are more emotionally expressive than men (Kring and Gordon, 1998). Participants were asked to imagine a situation in which his boss asked his impression about his favorite object. Twenty photographs were prepared as the objects to be praised, and 10 were photos of rooms. Some rooms were cool, clean, and looked comfortable because they were expected to arouse positive emotion. The others were of dirty and messy rooms because they were expected to arouse negative emotion. The remaining 10 photos were of a person who in different attire. Some clothes were fashionable and cool and were expected to arouse positive emotion. Others were eccentric and unstylish and were expected to arouse negative emotion. All the photographs were manually collected from the Internet.
Two types of compliment expressions were collected; praised with the participants' own expressions (free) and by uttering a fixed sentence, i.e. "綺麗な部屋ですね *(This is a very nice room)*" or "かっこいいですね *(You are so fashionable)*." With this process, 140 compliment videos were collected, and the utterances were about 2-7 seconds long. The resolution and the frame rate of the videos were 1,920x1,080 and 30fps, respectively. The bit depth and the sampling rate of the audio data were 16bit and 48,000Hz, respectively.
The compliment videos were annotated based on whether the speaker was sincere just after uttering the compliment. The participants annotated their compliments by choosing either {Sincere, Insincere, Neither, or Not_sure}. Table 5 lists the annotation results. Compliment videos annotated as *Sincere* and *Insincere* were only used for the experiment. The videos that annotated as *Neither* or *Not_sure* were excluded from experimental data because nobody, even the actual speaker itself, cannot define the correct answers of the data.

3.2. Experimental Results

The proposed method was evaluated using the leave-one-out cross validation method. Disproportion between sincere and insincere labeled data was solved by giving disproportionate weight when the SVM was applied.
Tables 6 and 7 list the SVM classification results by using each feature set (FACE: facial-expression-transition feature, ACO: acoustic feature, F+A) for free and fixed sentences, respectively. The results were compared with the random output and the annotation results by the others. The results of random output were calculated from the average in 100 trials. Meanwhile, the collected compliment videos were annotated by six participants (except the actual speaker of the utterance). All of participants were not expert lie catchers such as psychotherapists and police interrogators. When four or more participants annotated a video as *Sincere*, the annotation results were defined as *Sincere*, and vice versa. When three participants annotated as *Sincere* and the other three participants annotated as *Insincere*, the annotation results were defined as N/A. 27 of free and fixed sentence videos were annotated as N/A in this experiment, respectively.

Compliment	Annotation by actual speaker			
	Sincere	Insincere	Neither	Not_sure
Free sentence	80	45	7	8
Fixed sentence	63	64	8	5

Table 5: Number of data annotated by actual speaker

Feature set	Sincere			Insincere			accuracy
	precision	recall	F-measure	precision	recall	F-measure	
FACE	0.71	0.70	0.70	0.48	**0.49**	**0.48**	0.62
ACO	0.66	0.71	0.69	0.41	0.36	0.38	0.58
F+A	0.71	**0.75**	**0.73**	**0.50**	0.44	0.47	**0.64**
random	0.51	0.64	0.57	**0.50**	0.36	0.41	0.50
by human	**0.77**	0.58	0.66	0.45	0.38	0.41	0.50

Table 6: Experimental results (free sentence)

Feature set	Sincere			Insincere			accuracy
	precision	recall	F-measure	precision	recall	F-measure	
FACE	0.56	**0.60**	**0.58**	**0.58**	**0.53**	**0.55**	**0.57**
ACO	0.52	0.51	0.51	0.52	**0.53**	0.53	0.52
F+A	0.52	0.51	0.51	0.52	**0.53**	0.53	0.52
random	0.50	0.49	0.49	0.50	0.50	0.50	0.50
by human	**0.59**	0.52	0.55	0.57	0.39	0.46	0.46

Table 7: Experimental results (fixed sentence)

Table 6 indicates that the FACE feature set was the most superior among all of the classification methods to detect insincere compliments (F-measure (FACE)=0.48). On the contrary, the F+A feature set was the most superior to detect sincere compliments (F-measure(F+A)=0.73). Totally, the accuracy of the classifier using FACE feature set was improved by adding the ACO feature set (accuracy(F+A) > accuracy(FACE)). The results of the F+A feature set were superior than that of random output and the annotation result by the others as a whole.
Table 7 indicates that the FACE feature set was the most superior among all of the classification methods. However, the results using fixed sentences were worse than that using free sentences. One of the reasons is that using fixed sentences may reduce arousing negative emotions.

3.3. Features Correlated with Insincerity

The correlation ratios with insincerity were calculated for all features in the F+A feature set for free-sentence compliment data. Table 8 lists the rank of the features in which the correlation ratios were high. The names of features in FACE consists of *emotion*, *period*, and *feature*, and 1, 2, and 3 for *period* indicate the first, second, and third period of a contour divided into three parts, respectively. "All" indicates the whole contour.
In FACE features, *sad* and *anger*, especially *sad*, often correlate with insincere compliment. It indicates that micro/suppressed sad expression will appear when a

Rank	Feature set	Feature name
1	ACO	mfcc[4]_minPos
2	FACE	anger_2_range
3	ACO	pcm_zcr_delta_min
4	FACE	sad_2_range
5	ACO	voiceProb_delta_kurtosis
6	FACE	sad_2_min
7	ACO	mfcc[7]_delta_amean
8	ACO	mfcc[7]_delta_maxPos
9	ACO	mfcc[1]_max
10	ACO	voiceProb_delta_linregc1
11	FACE	sad_all_peak_maxPos(relative)
12	ACO	mfcc[3]_kurtosis
13	FACE	sad_1_range
14	ACO	mfcc[8]_kurtosis
15	ACO	mfcc_[2]_max

Table 8: High-correlation features (free sentence)

person praises something against his/her will. *Range in a period* also correlates with insincerely. It indicates that *range in a period* is valid to detect very short change of facial expression like micro expression. In ACO features, *zcr* and *voiceProb* often correlate with insincerity. These features relates short pause in an utterance.

3.4. Discussion

The experimental results in Section 3.2 showed that the classifier using F+A is superior than that using only FACE. It indicates that considering multimodal information worked better than single-modal information. However, present accuracy rate for detecting insincerity is not enough. To develop the system, further acoustic features should be added to consider speech rate, pause, voice quality, and so on. Facial-expression-transition features can be also improved. For example, dividing contours into shorter periods will support to detect much more clues of micro expressions.

Although our proposed method performed better than annotation results by non-expert participants, we have to compare the experimental results with annotation results by expert lie catchers, too.

4. Conclusion

We proposed a method of detecting concealed/suppressed negative emotion by using the transitions in facial expressions and acoustic features. To detect micro and squelched expressions, or abrupt onset/offset of facial expression, facial images for every 100 milliseconds were analyzed and the transition in the likelihood ratios for five emotions (happiness, sadness, surprise, anger, and neutral) were converted into contours. Then, each emotion contour was equally divided into three periods. A total of 90 features were calculated for these 20 contours. Also, 384 acoustic features related to pitch and power were calculated.

The experimental results indicate that the feature set including both of facial-expression-transition and voice was the most superior. Its precision and recall of insincerity detection and the total accuracy rate were 0.50, 0.44, and 0.64, respectively. The results were better than the annotation results by non-expert participants.

For future work, more effective acoustic features (e.g. features that correlate speech rate, pause, and voice quality) should be added to increase the accuracy rate of insincerity detection.

Acknowledgements

This research is supported by the Center of Innovation Program from Japan Science and Technology Agency, JST.

References

DeVault, D., Arstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G, Marsella, S.C., Fabrizio, M., Nazarian, A, Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A. and Morency, L.-P. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In: *Proceedings of AAMAS,* pp. 1061-1068.

Ekman, P. (2009). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition).* W.W. Norton.

Eyben, F., Wöllmer, M. and Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. In: *Proceedings of the International Conference on Multimedia*, pp. 1459-1462.

Hancock, J.T., Curry, L.E., Goorha, S. and Woodworth, M. (2008). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. In: *Discourse Processes*, 45(1), pp.1-23.

Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Shriberg, E. and Stolcke, A. (2005). Distinguishing Deceptive from Non-Deceptive Speech. In: *Proceedings of INTERSPEECH 2005*, pp. 1833-1836.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. In: *Journal of Computational and Graphical Statistics*, 5, pp. 299–314.

Kring, A.M. and Gordon, A.H. (1998). Sex differences in Emotion: Expression, Experience, and Physiology. In: *Journal of Personality and Social Psychology*, 74(3), pp. 686-703.

Li, Y., Inoue, K., Nakamura, S, Takanashi, K., Ishi, C.T. and Kawahara, T. (2017). Emotion Recognition by Combining Prosody with Text Information and Assessment Selection for Human-Robot Interaction. In: *Proceedings of JSAI SIG-SLUD-B506-09*, B5(3), pp. 43-48.

Lifedata (2015). *Praise Expression List.* Retrieved from: https://lifedata1.com/praise-summary-041. Access date: October 23, 2017. (in Japanese).

Mera, K., Ichimura, T., Kurosawa, Y. and Takezawa T. (2010). Mood Calculating Method for Speech Interface Agent by Using Emotion Generating Calculation Method and Mental State Transition Network. In: *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 22(1), pp. 10-24.

OMRON (2014). *OMRON's Image Sensing Site: +SENSING.* Retrieved from: https://plus-

sensing.omron.com/technology. Access date: October 23, 2017.

Regoulot, S., Fish, K. and Pell, M.D. (2014). Neural correlates of inferring speaker sincerity from white lies: An event-related potential source localization study. In: *Brain Research*, 1565, pp. 48-62.

Takahashi, T., Mera, K., Nhat, T.B., Kurosawa, Y. and Takezawa, T. (2017). Natural Language Dialog System Considering Speaker's Emotion Calculated from Acoustic Features. In: Jokinen, K. and Wilcock, G. (Eds) *Dialogues with Social Robots*, pp. 145-157, Springer Singapore.

Zhang, Z., Singh, V, Slowe, T.E., Tulyakov, S. and Govindaraju, V. (2007). Real-time Automatic Deceit Detection from Involuntary Facial Expressions. In: *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*.