Emotion Recognition of Conversational Affective Speech Using Temporal Course Modeling

Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan
{jenchunlin, chunghsienwu, lilijinjin}@gmail.com

Abstract

In a natural conversation, a complete emotional expression is typically composed of a complex temporal course representing temporal phases of onset, apex, and offset. In this study, sub-emotional states are defined to model the temporal course of an emotional expression in natural conversation. Hidden Markov Models (HMMs) are adopted to characterize the sub-emotional states; each represents one temporal phase. A sub-emotion language model, which considers the temporal transition between sub-emotional states (HMMs), is further constructed to provide a constraint on allowable temporal structures to determine an optimal emotional state. Experimental results show that the proposed approach yielded satisfactory results on the MHMC conversation-based affective speech corpus, and confirmed that considering the complex temporal structure in natural conversation is useful for improving the emotion recognition performance from speech.

Index Terms: Temporal course, hidden Markov model, emotion recognition

10.21437/Interspeech.2013-354

1. Introduction

Emotions play an important role in human intelligence, rational decision making, social interaction, perception, and memory [1]. Understanding latent meaning of affective speech is indispensable for day-to-day functioning of humans. With the growing and varied uses of human-computer interactions, emotion recognition technology has been used to provide harmonious interactions or communication between computers and humans [1-12]. Hence, constructing a high-performance emotion perception and recognition system from speech signal is highly desirable.

Although various studies in emotion recognition from speech have shown the benefits using different features and classifiers [2], [10], toward high-performance emotion recognition, an important issue is the dynamic aspects of emotional expression available for model training. In general, when the temporal course of emotional expression is complex, the temporal information could be lost owing to inappropriate model structures and this may lead to inaccurate estimates of the statistical model parameters; therefore, the classification result will be unsatisfactory. Previous research [1], [13-18], showed that a complete emotional expression can be characterized in three sequential temporal phases: onset (application), apex (release), and offset (relaxation), when considering the manner and intensity of an expression. To capture the temporal information, many design issues regarding the structure and the training process of the hidden Markov model (HMM) have been investigated.

In most HMM-based emotion recognition schemes, the left-to-right topology of the HMM structure was used [4], [19-21], and has been proven useful in modeling the signal streams (i.e., audio or visual) for describing the temporal courses of emotional expressions. However, it may be invalid for utterance-based emotion recognition, especially in natural conversation. Typically, a complete emotional expression is expressed by more than one utterance in natural conversation, and in more detail, each utterance may contain several temporal phases of emotional expression. Accordingly, a single HMM with left-to-right topology is unable to model the temporal course of emotional expression in natural conversation effectively. Figure 1 shows that when the emotional state (i.e., happiness) of Speaker 1 is evoked through conversation, Utterance 1 only covers the temporal phase of onset, while the apex and offset phases are covered in Utterance 2. For single HMM-based model training, the temporal information is lost by diverse training samples with complex temporal structures. Thus, for each emotional state, if a single HMM with left-to-right topology is used to model all the temporal phases of onset, apex, and offset, the performance may be degraded for emotion recognition of multiple utterances with complex temporal structures in natural conversation. An effective emotion recognition scheme in a real conversational environment is desirable to model the complex temporal structure.

To deal with this problem, in this study, we mainly focus on modeling the temporal expression evolution of an emotional state in an isolated sentence. Each isolated sentence in a conversation can express one or several sub-emotional states, which are defined to represent the temporal phases (i.e., onset, apex, or offset with high or low intensity) of an emotional expression. In this study, an HMM is used to characterize one sub-emotional state, instead of the entire emotional state. By integrating the sub-emotion language model, which model the temporal transition between sub-emotional states expressed in an isolated sentence, the proposed temporal course modeling scheme can further provide a constraint on allowable temporal structures to obtain an optimal recognition result of emotional state in each utterance. Accordingly, different from the current recognition schemes [10], in this study, we endeavor to solve the problem of complex temporal structures of emotional expressions for speech emotion recognition in a conversational environment.

The rest of the paper is organized as follows. Section 2 briefly outlines the procedure for feature extraction. Section 3 details the derivation of the proposed temporal course modeling approach. Section 4 shows the experimental results. Section 5 offers a conclusion.

2. Feature Extraction

An important issue for emotion recognition from speech is the selection of relevant features to be used. Hence, several promising features such as prosodic and acoustic features of affective speech signals have been discussed over the years.

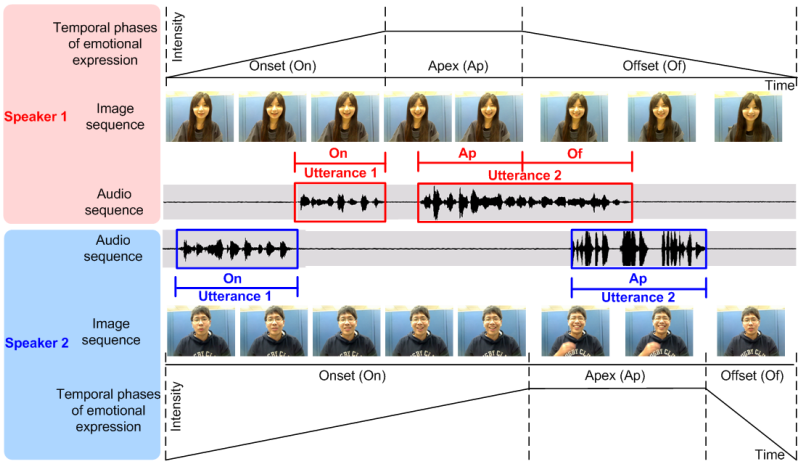
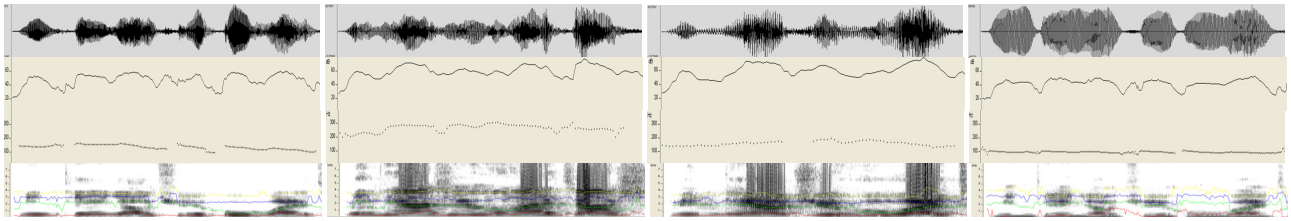


Figure 1: *An example of various temporal phases of happy emotional expression occurred to different utterances in a real conversational environment.*



(a) (b) (c) (d)

Figure 2: *An example for illustrating the difference of prosodic features (from top to bottom: waveform, energy, pitch, and formants) among (a) neutral, (b) happy, (c) angry, and (d) sad emotional states.*

Among these features, in speech emotion recognition [2], [5] [6], [22-24] prosodic features have been found to represent the most significant characteristics of emotional content in verbal communication. Several studies [2], [25] have further noted that pitch and energy are useful to determine emotion in speech. In addition, Morrison et al. [6] further summarized the correlations between prosodic features and emotions. These findings concluded that prosody-related features are highly beneficial to emotion recognition.

To verify this argument, the prosodic features of a Chinese sentence with four emotional states are shown in Figure 2. According to our observations from the energy contour, happy and angry emotional states have higher intensities compared to sad and neutral states. In pitch contour, the pitch ranges and pitch levels of sad emotion are narrower and lower than those of other emotional states. Figure 2 highlights the differences in the prosodic features of the four emotional states. This observation strengthened the analyses of the mentioned research. In addition, formants and speaking rate were also frequently employed and discussed [6], [8], [10]. Although speaking rate is an important prosodic feature for emotion recognition, incorrect detection of the speaking rate resulting from speech recognition error (especially in affective speech) will dramatically degrade the emotion recognition performance. In this study, three types of primary prosodic features, including pitch, energy and formants F1-F5 in each speech frame, were used for emotion recognition.

For prosodic feature extraction, the pitch detection tool "Praat" was used [26]. Normalization of the extracted prosodic features is needed because of the person-to-person variation and recording conditions. Thus, for each subject, the prosodic features of every frame are normalized by their means individually from the neutral expression sequence. The neutral sequence was manually selected from the video sequences at the beginning.

3. Temporal Course Modeling

For complex temporal structure modeling, in this study, since an emotional state may cover one to several sentences, a sequence of M temporal phases, each modeled by a sub-emotion HMM, is used to characterize an emotional state expressed in an isolated sentence. In order to better describe the temporal course of an emotional expression, a sub-emotion language model similar to the language model widely and successfully used in speech recognition is employed. By integrating the sub-emotion language model, the proposed temporal course modeling scheme can further provide a constraint on allowable temporal phase sequences to determine an optimal emotional state in an isolated utterance.

For model derivation, the recognition task with four emotional states, happy, angry, sad, and neutral, represented by $E \in \{H, A, S, N\}$, is considered. Since humans are likely to express emotions with various intensities (e.g., we can be (feel) more or less happy), to model the temporal course of emotional expression precisely, it is inevitable to integrate the intensity information into temporal phases. Figure 3 shows an example to describe an emotional expression which can have different expressivity of intensity across time. Accordingly, each emotional state of happy, angry, and sad for each isolated sentence is further expressed by an M-temporal phase sequence, denoted as $H = h_1^M = h_1, h_2, ..., h_M$, $A = a_1^M = a_1, a_2, ..., a_M$ and $S = s_1^M = s_1, s_2, ..., s_M$, where M ranges from 1 to 5 considering all possible transitions in a predefined grammar shown in Figure 4. Each sub-emotion HMM in the sequence is directly used to model a temporal

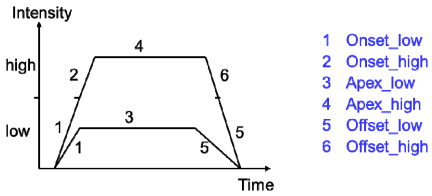


Figure 3: *An example of the emotional expression for temporal phases with different intensities.*

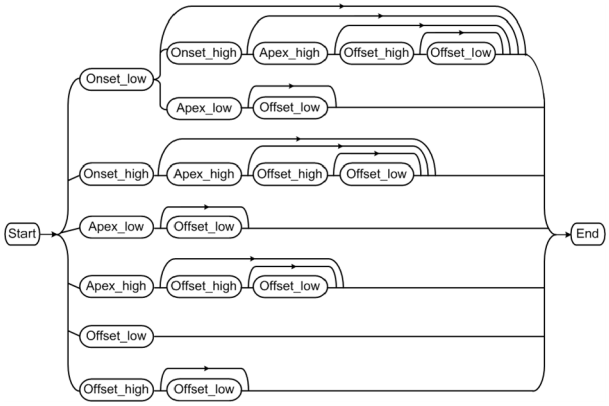


Figure 4: *Recognition network based on the predefined grammar for characterizing an emotional state expressed in an isolated sentence.*

phase. Although the neutral vocal expressions may evolve over time, the temporal phases of neutral emotional state are unobvious and difficult to label manually. In this study, we simplified the expression evolution of the neutral emotional state as only one neutral temporal phase. Given the observation sequence $O = o_1^T = o_1, o_2, ..., o_T$, the probability of an emotional state with temporal phase sequence E can be estimated using (1), where $\hat{E}$ represents the emotion recognition result by maximizing the *a posteriori* probability $P(E \mid O)$.

$$\hat{E} = \arg\max_{E} P(E \mid O) \tag{1}$$

The *a posteriori* probability $P(E \mid O)$ can be further decomposed using the Bayes' rule as follows:

$$P(E \mid O) = P(O \mid E)P(E)/P(O) \tag{2}$$

where $P(O \mid E)$, denoting the likelihood of the observation, is calculated using the corresponding sub-emotion HMM sequence, and $P(E) = P(e_1, e_2, ..., e_M)$ is the *a priori* probability of observing temporal phase sequence $E = e_1, e_2, ..., e_M$ and is estimated by the sub-emotion language model. A bigram language model is adopted and constructed to estimate the probability $P(E) = P(e_1, e_2, ..., e_M) = \prod_{k=2}^{M} P(e_k \mid e_{k-1})$. A recognition network for the pre-defined grammar shown in Figure 4 is constructed based on the temporal phase definition [13-16]. $P(O)$ in (2) is the same for all possible recognized emotional states with temporal phase sequence E, and can thus be omitted. Hence, (1) can be rewritten as (3) for emotion recognition using the proposed temporal course modeling.

Table 1. *The ground truth utterances of four emotional states for MHMC conversation-based affective speech corpus.*

	Happy	Angry	Sad	Neutral
# (Utterances)	199	236	214	465

Table 2. *Average emotion recognition rates of four emotional states in different models.*

Models	SVM	Traditional HMM	Proposed
Accuracy	50.22%	56.50%	79.82%

Table 3. *The confusion matrix of the proposed temporal course modeling approach for four emotional states.*

	Happy	Angry	Sad	Neutral	Accuracy
Happy	31	1	2	0	91.18%
Angry	18	25	1	1	55.56%
Sad	4	0	32	2	84.21%
Neutral	6	7	3	90	84.91%

$$\hat{E} = \arg\max_{E} P(O \mid E)P(E) \tag{3}$$

In the training procedure, the proposed temporal course modeling consists of three parts:

- Given the labeled training data, the parameter sets of sub-emotion HMMs were trained using the expectation-maximization (EM) algorithm.
- The optimal state sequence was obtained using the Viterbi algorithm in the corresponding sub-emotion HMM sequences.
- The sub-emotion language model $P(E)$ was constructed by estimating the temporal phase transitions according to the labeled temporal phase sequences in the training data using the bigram language model.

4. Experimental Results

This study evaluated the performance of the proposed method based on an affective speech corpus. The conversation-based affective speech corpus [18], collected from Multimedia Human-Machine Communication (MHMC) Laboratory, was provided by 53 students of both genders in National Cheng Kung University, Taiwan. During the recording session, toward naturalistic conversation, a conversation topic was first selected by each paired participants, and for each topic, the participants spoke as they like instead of navigating a pre-design script. For four emotional states, a total of 2,120 utterances were collected to form the MHMC conversation-based affective speech corpus.

The subjective tests were performed to set the ground truth of emotional expression for the recorded data. Hence, three annotators were recruited from the MHMC laboratory, and each of them was asked to give an opinion on the emotion label for the recorded data. During the labeling process, the annotators were allowed to check the vocal expression of the recorded data more than once to ensure that the labels can truly reflect their feelings. After the labeling process, each labeled data was then evaluated by checking the opinions of all annotators. If less than two annotators reached an agreement, the data was not included in the experiment. Regarding the labels of the temporal courses of emotional expressions, same as the process of emotion labeling, the

subjective tests were performed to set the ground truth of the temporal phases for the recorded data. Before labeling, we have explained the temporal phase definition and provided four to six examples to show the emotional expressions covering various temporal phases to each annotator. Accordingly, after the labeling process, the ground truth was decided when the labeled temporal phase categories from at least two annotators are the same. Finally, a total of 1,114 data, which passed the evaluation (i.e., simultaneously passed the emotion and temporal phase labeling procedure), were regarded as the ground truth data for the ensuing experiments. The number of ground truth utterances of the four emotional states is shown in Table 1.

In the experiments, two popular classifiers were considered: Support Vector Machine (SVM) and HMM [10], [19], [21]. For performance evaluation, the SVM with radial basis kernel function and the left-to-right topology of the HMM structure with eight hidden states (i.e., achieved the best recognition accuracy) were used for comparison. For SVM, the global features [10] were used in which the minimum, mean, and maximum of the extracted prosodic features were considered. In terms of the proposed approach, the left-to-right topology of the HMM structure with three hidden states was applied for modeling each temporal phase. In the experiments, 80% of the ground truth utterances were randomly selected from the MHMC conversation-based affective speech corpus for training, and the remaining utterances were selected for testing.

The average recognition accuracy for three approaches is shown in Table 2. The confusion matrix of the proposed approach is further shown in Table 3. The results in Table 2 show that comparing with SVM and the traditional HMM approaches [10], [19], [21], the proposed approach of the temporal course modeling achieved the best recognition accuracy. The results confirmed that considering the temporal phases and combining the sub-emotion language model is able to better describe the complex temporal structure of emotional expression in natural conversation. In addition, the proposed temporal course modeling scheme considering expression intensity (i.e., low or high intensity for each temporal phase) is helpful for differentiating the expression styles between introvert and extrovert. Because the emotional expressions of the introverts are quite different from those of extroverts (e.g., since the introvert is often bashful, the expressed emotions are often accompanied with lower expression intensity), the diverse expression styles may lead to large variations in the statistical model parameters, and thus degrade recognition performance. Compared to the previous approaches, with the property of expression intensity, the proposed temporal course modeling is helpful for alleviating the effect of diverse expression styles which can diminish the variations of model parameters and distinguish the expression styles. The Chi-squared test statistic further shows that the differences of performances over three models were statistically significant with $\chi^2(2) = 46.122, P < .0001$. In addition, using post hoc pair-wise comparisons for Chi-squared test, the crosstab (also known as Contingency Table) based on the z-test with Bonferroni corrections is shown in Table 4 [27], [28]. The results show that the correct recognition results of the proposed method (C) are statistically more significant than those of other models (i.e., A, B). The statistical analysis verified the effectiveness of the proposed method.

Table 4. *A crosstab of post hoc pair-wise comparisons for Chi-squared test based on the z-test with Bonferroni corrections.*

Significance Comparison	MODEL		
	SVM (A)	Traditional HMM (B)	Proposed (C)
Correct			AB
Error	C	C	

Table 5. *Average emotion recognition rates of high and low arousal categories in different models.*

Models	SVM	Traditional HMM	Proposed
Accuracy	85.20%	89.69%	89.69%

In addition, the proposed method can also reduce the misclassification of the similar arousal characteristic of emotions such as happy and angry emotional states. To further verify our contention, the performance of arousal classification experiments is shown in Table 5; happy and angry states are regarded as the relatively high arousal emotions, and the neutral and sad states are regarded as relatively low arousal emotions which form two arousal categories. According to the observation, we find that the three approaches achieved comparable results for arousal classification (i.e., three approaches achieved approximately 90% average recognition accuracy). The experiments first strengthened the literature analyses [1], that is, vocal expressions are useful for arousal classification. Compared to the results in Table 2, the experiments also demonstrated that modeling the temporal information accompanied with the expression intensity is important to distinguish the emotions which contain the same characteristic of arousals. The results show that the performances of the SVM and traditional HMM approaches were dramatically degraded from 85.20% and 89.69% for arousal classification to the 50.22% and 56.50% for emotional states classification, respectively. Hence, the results reflected that missing the temporal and expression intensity information have significant effects on the recognition results of SVM and traditional HMM approaches for recognizing four emotional states in natural conversation. Based on these analyses, the findings revealed that decoding the complex temporal structure of emotional expression and considering expression styles are helpful to improve the recognition accuracy for conversation-based emotion recognition.

5. Conclusion

This paper presented an approach to automatic recognition of four emotional states and arousal categories from speech signal using HMM-based temporal course modeling. Two findings are summarized from our experiments. First, modeling the complex temporal structure of emotional expression is helpful for improving the recognition accuracy. Second, modeling the expression styles is useful for emotion recognition. The expression styles, such as expression intensity between introverts and extroverts, are significantly different for the same emotional state. Compared to previous approaches, the experiments demonstrated the proposed approach can provide better abilities for describing the diverse expression styles and complex temporal information for emotion recognition in natural conversation. For effective emotion recognition, future research to explore the expression styles from different users is a viable direction, which may be further related to the expression manner and significantly associated to the personality trait.

6. References

[1] Picard, R. W., Affective Computing. MIT Press, 1997.

[2] Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S., "A survey of affect recognition methods: audio, visual, and spontaneous expressions", IEEE Trans. Pattern Anal. Mach. Intell., 31(1):39-58, 2009.

[3] Wu, C. H. and Liang, W. B., "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels", IEEE Trans. Affective Computing, 2(1):1-12, 2011.

[4] Lin, J. C., Wu, C. H. and Wei, W. L., "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition", IEEE Trans. Multimedia, 14(1):142-156, 2012.

[5] Wu, C. H., Yeh, J. F. and Chuang, Z. J., "Emotion perception and recognition from speech", Affective Information Processing, Chapter 6, 93-110, Springer, 2009.

[6] Morrison, D., Wang, R. and De Silva, L. C., "Ensemble methods for spoken emotion recognition in call-centres", Speech Communication, 49(2):98-112, 2007.

[7] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G., "Emotion recognition in Human-Computer Interaction", IEEE Signal Processing Magazine, 33-80, 2001.

[8] Lee, C. M. and Narayanan, S., "Toward detecting emotions in spoken dialogs", IEEE Trans. Speech and Audio Processing, 13(2):293-303, 2005.

[9] Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D. and Levinson, S., "Audio-visual affect recognition", IEEE Trans. Multimedia, 9(2):424-428, 2007.

[10] Ayadi, M. E., Kamel, M. S. and Karray, F., "Survey on speech emotion recognition: features, classification, schemes, and databases", Pattern Recognition, 44(3):572-587, 2011.

[11] Wu, C. H., Chuang, Z. J. and Lin, Y.C., "Emotion recognition from text using semantic label and separable mixture model", ACM Trans. on Asian Language Information Processing, 5(2):165-183, 2006.

[12] Wu, C. H., Wei, W. L., Lin, J. C. and Lee, W. Y., "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion", to appear in IEEE Trans. Multimedia, 2013.

[13] Ekman, P., Handbook of Cognition and Emotion. Wiley, 1999.

[14] Mana, N. and Pianesi, F., "Modeling of emotional facial expressions during speech in synthetic talking heads using a hybrid approach", Int'l Conf. Auditory-Visual Speech Processing (AVSP), 2007.

[15] Valstar, M. F. and Pantic, M., "Fully automatic facial action unit detection and temporal analysis", Proc. Int'l Conf. on Computer Vision and Pattern Recognition (CVPR), 2006.

[16] Valstar, M. F. and Pantic, M., "Fully automatic recognition of the temporal phases of facial actions", IEEE Trans. Systems, Man and Cybernetics–Part B, 42(1):28-43, 2012.

[17] Wu, C. H., Lin, J. C. and Wei, W. L., "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course", to appear in IEEE Trans. Multimedia, 2013.

[18] Wei, W. L., Wu, C. H., Lin, J. C. and Li, H, "Interaction style detection based on cross-correlation model in spoken conversation", to appear in Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[19] Schuller, B., Rigoll, G. and Lang, M., "Hidden Markov model-based speech emotion recognition", Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), II 1-4, 2003.

[20] Song, M., You, M., Li, N. and Chen, C., "A robust multimodal approach for emotion recognition", Neurocomputing, 71(10-12):1913-1920, 2008.

[21] Ntalampiras, S. and Fakotakis. N., "Modeling the temporal evolution of acoustic parameters for speech emotion recognition", IEEE Trans. Affective Computing, 3(1):116-125, 2012.

[22] Scherer, K. R., "Vocal communication of emotion: a review of research paradigms", Speech Communication, 40(1-2):227-256, 2003.

[23] Luengo, I., Navas, E., Hernáez, I. and Sánchez, J., "Automatic emotion recognition using prosodic parameters", Proc. INTERSPEECH, 493-496, 2005.

[24] Kooladugi, S. G., Kumar, N. and Rao, K. S., "Speech emotion recognition using segmental level prosodic analysis", Int'l Conf. on Devices and Communications, 1-5, 2011.

[25] Kwon, O. W., Chan, K., Hao, J. and Lee, T. W., "Emotion recognition by speech signals", Proc. Eighth European Conf. Speech Comm. and Technology (EUROSPEECH), 2003.

[26] Boersma, P. and Weenink, D., Praat: doing phonetics by computer. http://www.praat.org/. 2007.

[27] Cooper, H. M., Hedges, L. V. and Valentine, J. C., The Handbook of Research Synthesis and Meta-Analysis. Russell Sage Foundation, NY 2009.

[28] Toothaker, L. E., Multiple Comparison Procedures. Sage Pubns, 1992.