AFRL-OSR-VA-TR-2013-0551

(YIP-10) COMPUTATIONAL MODELING OF EMOTIONS AND AFFECT IN SOCIAL-CULTURAL INTERACTION

YANG LIU

UNIVERSITY OF TEXAS AT DALLAS

10/02/2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30-09-2013	Final report	July 15, 2010-July 14, 2013

4. TITLE AND SUBTITLE

Computational Modeling of Emotions and Affect in Social-Cultural Interaction

5a. CONTRACT NUMBER

5b. GRANT NUMBER

AFOSR FA9550-10-1-0388

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Yang Liu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, Texas 75080

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
875 North Randolph Street, Suite 325, Room 3112,
Arlington, VA., 22203-1768

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A - Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of this proposal was to develop computational models for emotion recognition in speech and study various impacting factors including social, cultural, and language effect on such models. Accomplishments in the project are the following. First, emotion recognition performance was improved upon the state-of-the-art. Different methods were developed to improve model performance, including employing sub-sentence units, advanced feature transform, deep learning, and more features from acoustic and textual information sources. Second, a cross-lingual study was performed that shed light on how human perception and automatic recognition of emotion differs and how performance in cross-lingual setups varies. This project supported several students, leading partly to one Ph.d dissertation.

15. SUBJECT TERMS

emotion/affect recognition, social and cultural effect, cross lingual study

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UU		**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Grant Number: AFOSR FA9550-10-1-0388

Project Title: (YIP-10) Computational Modeling of Emotions and Affect in Social-Cultural Interaction

Start date of project: July 15, 2010

End date of project: July 14, 2013

Technical point of contact: Yang Liu

Final project report

Proposal Title: Computational Modeling of Emotions and Affect in Social-Cultural Interaction

Technical point of contact: Yang Liu, 972-883-6618, yang.liu@utdallas.edu

Project goals:

The objective of this project was to develop computational models for recognizing emotion in speech using a rich set of prosodic, acoustic, and lexical features that capture how speech is produced, a speaker's pitch and intonational pattern, and word usage. Better feature representation and advanced approaches were used to improve emotion recognition performance. Another goal is to conduct cross lingual and cultural analysis of affective behavior. This research represented a major advance in the state of the art in emotion recognition in speech/language. The outcomes of this project contributed to fundamental understanding of how emotions are signaled and how to successfully model this phenomenon, which has an impact on using speech technology in various applications.

Primary accomplishment and findings:

The research conducted as part of this project advanced the state-of-the-art automatic emotion recognition performance, and improved our understanding of language/cultural impact on human perception of emotion and automatic classification.

- Units used. This study investigates sentence-level emotion recognition. We proposed to use a two-step approach to leverage information from subsentence segments for sentence level decision. First we train a segment level emotion classifier, and generate predictions for segments within a sentence. A second component combines the predictions from these segments to obtain a sentence level decision. We evaluated different segment units (words, phrases, time-based segments) and different decision combination methods (majority vote, summation of probabilities, and a Gaussian mixture model). Our experimental results have shown that our proposed method significantly outperforms the standard sentence-based classification approach. In addition, we find that using time-based segments achieves the best performance, and thus no speech recognition or alignment is needed when using our method. This is important when extending emotion recognition to languages that do not have speech recognizers available.
- Level of interest detection. In this study, we proposed a decision-level fusion approach using acoustic and lexical information to accurately sense a user's interest at the utterance level. Our system consists of three parts: acoustic/prosodic model, lexical model, and a model that combines their decisions for the final output. We use two different regression algorithms to complement each other for the acoustic model. For lexical information, in addition to the bag-of-words model, we propose new features including a level-of-interest

value for each word, length information using the number of words, estimated speaking rate, silence in the utterance, and similarity with other utterances. We also investigate the effectiveness of using more automatic speech recognition (ASR) hypotheses (n-best lists) to extract lexical features. The outputs from the acoustic and lexical models are combined at the decision level. Our experiments show that combining acoustic evidence with lexical information improves level-of-interest detection performance, even when lexical features are extracted from ASR output with high word error rate.

- Better feature and models. We have investigated two different modeling approaches to improve features and models used for emotion recognition, based on i-vector models and deep learning methods, respectively.
    - Using i-vector space features has been shown to be very successful in speaker and language identification. In our study, we evaluated using the i-vector framework for emotion recognition from speech. Instead of using standard i-vector features, we proposed to use concatenated emotion specific i-vector features. For each emotion category, a Gaussian mixture model (GMM) supervector is generated via adaptation of the neutral one from a big corpus. An i-vector feature vector is then obtained using each emotion specific GMM supervector. The concatenation of these emotion dependent i-vector features is used as the feature vector in the backend emotion classifier, e.g., support vector machines (SVM). Our experimental results on acted and spontaneous data sets demonstrate that this method outperforms baselines using either static features or dynamic features.
    - Deep learning has been recently widely used in various machine learning problems, including tasks in speech and language processing. We investigated using the denoising autoencoder to generate robust feature representations for emotion recognition. In our method, the input of the denoising autoencoder is the normalized static feature set (state-of-the-art features for emotion recognition). This input is mapped to two hidden representations: one is to capture the neutral information from the input, and the other one is used to extract emotional information. Model parameters are learned by minimizing the squared error between the original and the reconstructed input. After pre-training and fine-tuning, we use the hidden representation as features in the SVM model for emotion classification. Our experimental results show significant performance improvement compared to using the static features.
- Cross language study. The aim of this study is to investigate the effect of cross-lingual data on human perception and automatic classification of emotion from speech. We use four different databases from three languages (English, Chinese, and German) and two types (acted and improvised). For automatic classification, there is a significant degradation using cross-corpus than within-corpus setup. For human perception, we observe differences between native and non-native speakers when judging emotions for a language, and there is less performance loss in cross-language setup compared to automatic classification. In addition, we find that the automatic approaches work well in classifying the emotional activation category: positive and negative activated emotions, but are not good at classifying instances within the same activation category, which is different from the confusion patterns of the human perception experiment. This study provides insights to better understanding of cross-lingual human emotion perception and development of robust automatic emotion recognition systems.

Staff supported under the grant:

Faculty:

Yang Liu

Students:

Main students supported by the project:

- Je Hun Jeon (Ph.d student, graduated, now working at Nuance)
- Rui Xia (Ph.d student)
- Duc Le (undergraduate student, graduated, now pursuing Ph.d at University of Michigan.)

Other students supported shortly by the project for exploratory studies related to this project:

- Dong Wang (Ph.d student, studied sentiment analysis in conversational speech)
- Khairun Nisa Hassanali (Ph.d student, studied opinion/emotion expression in child languages)

Grant related publications:

Journal papers:

1. Je Hun Jeon, Rui Xia, and Yang Liu. Level of Interest Sensing in Spoken Dialog Using Decision-level Fusion of Acoustic and Lexical Evidence. Computer Speech and Language. Accepted. 2013.
2. Je Hun Jeon and Yang Liu. Automatic Prosodic Event Detection Using Novel Labeling Method in Co-Training Algorithm. Speech Communication, Volume 54, Issue 3, March 2012, Pages 445-458.

Conference papers:

3. Rui Xia and Yang Liu. Using Denoising Autoencoder for Emotion Recognition. Interspeech, Lyon, France, 2013.
4. Je Hun Jeon, Duc Le, Rui Xia, and Yang Liu. A Preliminary Study of Cross-lingual Emotion Recognition from Speech: Automatic Classification versus Human Perception. Interspeech, Lyon, France, 2013.
5. Rui Xia and Yang Liu. Using i-Vector Space Model for Emotion Recognition. Interspeech, Portland, Oregon, 2012.
6. Je Hun Jeon, Rui Xia, and Yang Liu. Sentence Level Emotion Recognition based on Decisions from Subsentence Segments. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic, 2011.

7. Je Hun Jeon, Rui Xia, and Yang Liu. Level of Interest Sensing in Spoken Dialog Using Multi-level Fusion of Acoustic and Lexical Evidence. Interspeech, Makuhari, Japan, 2010.

Abstract:

The objective of this proposal was to develop computational models for emotion recognition in speech and study various impacting factors including social, cultural, and language effect on such models. Accomplishments in the project are the following. First, emotion recognition performance was improved upon the state-of-the-art using different methods. (I) Emotion predictions for sub-sentence segments are aggregated to form the final decision for the sentence. (II) Acoustic prosodic cues and textual information are combined at different levels to decide the final emotion for an utterance. In addition, multiple speech recognition outputs are used to extract textual features, instead of just one single recognition hypothesis. (III) Advanced feature transform methods are employed to obtain more robust feature representation for emotion classification. Second, a cross-lingual study was performed that shed light on how human perception and automatic recognition of emotion differs and how performance in cross-lingual setups varies. This project supported several students, leading partly to one Ph.d dissertation.