The Social Sciences 11 (16): 3828-3832, 2016
ISSN: 1818-5800

Universality and Diversity of Cultural-Influenced Speech Emotion Recognition System

[1]Norhaslinda Kamaruddin, [2]Abdul Wahab, [1]Muhammad Jaliluddin Mazlan and [1]Norul Ayny Norzilan
[1]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi Mara Melaka, Jasin Campus, 77300 Merlimau, Melaka, Malaysia
[2]Kulliyah of Information Technology and Communication,
International Islamic University Malaysia (IIUM), 53100 Jalan Gombak,
Kuala Lumpur, Malaysia

Abstract: Culture refers to the cumulative knowledge, beliefs, values and concepts that are accepted by a group of people. Such information are shared and inherited from the previous generations in order for one to be blended and accepted in a society. Different cultural groups communicate differently that is distinct and unique making homogeneous interpretation of underlying emotional contents are more accurate. However, universality of cultural-influenced speech can be observed when cross cultural speeches are being interacted from different cultural groups to one another especially with the advancement of communication technology. In this study, two different cultural-influenced speech datasets representing American (NTU-American) and European (Netherland EmoSpeech) are employed to investigate their similarity and dissimilarity in term of heterogeneous listener's perception on the underlying emotional contents. The Mel Frequency Cepstral Coefficient (MFCC) feature extraction method and Multi Layer Perceptron (MLP) classifier are coupled to determine four different emotions, namely; anger, happiness, sadness and neutral acting as emotionless state. From the experimental result, it is noted that the proposed approach yielded accuracy performance of two times better than chance guessing. Moreover, the Netherland EmoSpeech dataset managed to obtain comparative accuracy with the established NTU-American dataset demonstrating that the data is satisfactory for speech emotion recognition purposes.

Key words: Cultural universality, cultural diversity, cultural-influenced speech emotion recognition, MFCC, MLP

INTRODUCTION

Speech is one of the simplest method of communication that comprises a two-way process of reaching mutual understanding between the speaker and the audience. The message is encoded by the speaker and transmitted through communication channels prior to thedecoding process by the listener. The listener will then infer the intended message and typically responds with an appropriate feedback. The naturalness of such interaction is the basis for researchers to emulate for a better Human Computer Interaction (HCI) in such a way that a computer has the ability to recognize emotion in a similar way to recognition. An emotion may influence the way an individual speak at a specific time. The speech prosodic features such as rhythm, intonation and stress as well as the volume, speaking rate and pitch are dynamically changed when the speaker expressing different emotions. For instance, the word 'please' may carry different meanings if it is articulated differently. If it is uttered in a higher pitch, it may represented disgust whereas if it is uttered with slow pitch, it may be perceived as a persuasive. Hence emotion can be identified and measured using speech. In this study, four different emotions, namely; anger, happiness, sadness and neutral are exclusively studied to give some insight from the perspective of culture universality and diversity.

Speech emotion universality and diversity: The expression of emotion is governed by the social norm and have significant impact on the intensity of emotion expression (Ekman, 1971). For instance, American uses communication style that is more linear, direct, detached, intellectually engaged and concrete while expressing anger while in contrast Asian prefers circular, indirect, attached and relationally engaged communication style

Corresponding Author: Norhaslinda Kamaruddin, Faculty of Computer and Mathematical Sciences,
Universiti Teknologi Mara Melaka, Jasin Campus, 77300 Merlimau, Melaka, Malaysia

(Matsumoto, 2001). This is because Americans are more focused on the individual happiness (individualistic culture) compared to Asian who relies on collective group satisfaction such as group conformity (collectivistic culture) (Miyamoto *et al*., 2010). Due to this notion, Asian often associates anger with rudeness which will affect the overall community feeling when overtly expressed. Thus, it is no surprise to state that culturedoes influence the conveyance and perception of emotion. There are common tendencies in the acoustical correlates of basic emotions across different cultures. The universality of the acoustic features allow people with widely varying cultural background to recognize emotion conveyed by other speaker regardless of the language or geographical boundary. This notion is supported by the result of the accuracy level greater than predicted by chance guessing (Elfenbein and Ambady, 2002; Scherer *et al*., 2001). Yet, there is a clear distinction of the way emotion conveyed and perceived by each culture such that these information is only shared among the people in the same culture (Dewaele, 2008). Beaupre and Hess (2005) mentioned that the reason for individual to accurately recognize expressions by members of their own ethnic group is because of their ability in encoding and decoding the expressions. For instance, there might be subtle differences in expressive style shared between similar members of a particular cultural group that make it more difficult to decode by other different cultural group member. Therefore, it is imperative that both intra-cultural and inter-cultural variation must be investigated in order to understand the cultural effect on speech emotion. Intra-cultural assessment is a similarity measure that enable different emotions to be detected using the same set of cultural parameters. For instance, given a similar culturally influenced dataset, human are able to discriminate anger, happiness and sadness (example of different emotion). On the contrary, inter-cultural assessment is a dissimilarity measure that enable a similar emotion (for example anger) to be detected across the different culturally influenced datasets. Homogeneous identification is conducted when one from the same culturally influenced group categorizes the emotion from his/her group. For instance, when an Asian labels the emotion from the speech emotion dataset that is influenced by Asian culture. In contrast, heterogeneous identification is conducted when an individual from different cultures label another cultural speech emotion dataset. For example, when an Asian categorized American speech emotion dataset, it is known as heterogeneous speech emotion identification. In this study, heterogeneous intra- cultural assessment is studied where the participants are Asian who identify American (NTU-American) (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2009; Kamaruddin and Wahab, 2012) and European (Netherlands EmoSpeech) culturally influenced datasets. NTU-American dataset is selected because based from previous experiments, it yielded the best accuracy performance (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2009; Kamaruddin and Wahab, 2012a).

Data collection and heterogeneous human listening survey: Datasets used in this project are NTU-American (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2009a; Kamaruddin and Wahab, 2012) and Netherland EmoSpeech datasets representing American and European cultural influence, respectively. NTU-American dataset results and analyses has been published in previous works (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2009; Kamaruddin and Wahab, 2012) and will be used as a comparative basis for the newly collected Netherlands EmoSpeech dataset. The Netherlands EmoSpeech dataset is uttered by four speaker (1 male and 3 female) using Dutch language with age ranging between 25-31 years old (mean 28.5 years and standard deviation of 2.5 years). The speakers are not trained professional actor and reported to experience calm emotion while their speech data are collected. The speakers are asked to speak the prepared text in their interpretation of four different emotions, namely; anger, happiness, sadness and neutral to represents emotionless state. Each emotion speech data (16 files) is recorded in wma format using 44.1 KHz sampling rate. The data then are down sampled to 8000 Hz and reformatted to wav file to mimic telephony system. This is because human can still identify different emotions expressed using telephone conversation. The heterogeneous human listening survey is conducted to verify the label of speech emotion from the prior data collection. It is based on the emotion perceived by the judges. Forty Asian human judges (N = 40) with age ranging between 19 and 24 participated in the study (mean age is 21.3 years old). Most judges are the undergraduate students of Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Shah Alam who does not understand Dutch language at all. The reason is to ensure that the speech emotion categorization is based on the acoustic information that the judges perceived rather than the semantic understanding of the words (Miyamoto *et al*., 2010). The heterogeneous human judges group are comprised of equal gender distribution (20 male and 20 female judges). Such arrangement is prepared to guarantee that gender does not influence the deduction of such observation. The data collected

Table 1: Homogeneous human listening survey result

File	A	H	N	S	Other
S1-A	85	5.0	0.0	2.5	7.5
S1-H	25	67.5	0.0	0.0	7.5
S1-N	0	0.0	72.5	27.5	0.0
S1-S	0	0.0	0.0	100	0.0
S2-A	95	5.0	0.0	0.0	0.0
S2-H	5	87.5	7.5	0.0	0.0
S2-N	0	0.0	95	5.0	0.0
S2-S	0	0.0	0.0	100	0.0
S3-A	85	7.5	5.0	0.0	2.5
S3-H	0	82.5	15.0	0.0	2.5
S3-N	5	0.0	82.5	2.5	10.0
S3-S	0	0.0	0.0	12.5	87.5
S4-A	62.5	20.0	0.0	10.0	7.5
S4-H	2.5	97.5	0.0	0.0	0.0
S4-N	0	0.0	72.5	22.5	5.0
S4-S	0	2.5	15.0	75.0	7.5

previously are randomly arranged in such a way that the judges cannot predict the sequence of speech emotion data presented. The judges are comfortably seated in front of a computer with a headphone in a controlled lab environment to minimize interruptions. Before the heterogeneous user listening survey is conducted, each judge needs to complete the demographic information questionnaire. In this study, only judges under no influence of medicine or drug and experiencing calm emotion are considered. Such extent is vital to ensure the results are not bias. The heterogeneous human listening survey usually took around 5-10 min where the judges are asked to carefully listen to the data collected and select the emotion that they perceived is being uttered by the speaker. The result are then compared to the actual speech emotion label and summarized in Table 1. The S is representing the speaker while A, H, N and S represent anger, happiness, neutral and sadness respectively. From the result in Table 1, judges always misjudged anger and happiness that show quite high misclassification between this two emotions. Such result may be because both anger and happiness are accompanied with high tone and fast speaking rate. Similar pattern is also observed between neutral and sad where judges always mistakenly perceived due to their lower pitch and slower speaking rate. It is noted that human from different culture (judges are Asian) managed to categorize the different emotion from different cultural-influenced speech emotion dataset. It demonstrates that the acoustical information is sufficient to correctly classify different emotions.

MATERIALS AND METHODS

Feature extraction method, classifier and experimental set-up: Once the data are verified, relevant features are extracted. In this study, Slaney's Mel Frequency Cepstral Coefficient (MFCC) feature extraction method is adopted

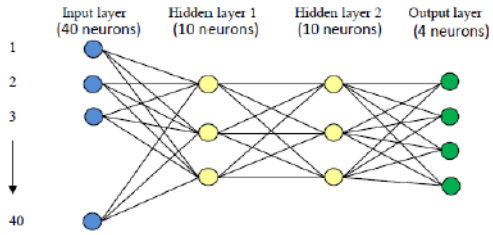

Fig. 1: MLP architecture

(Slaney, 1998). MFCC is based on the human hearing perceptions that mimic cochlear. It is one of the most widely used spectral features with a simple calculation, good ability of the distinction and high robustness to noise (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2009; Al-Ayadi *et al*., 2011). The sampling rate value used in this project is 8000 Hz and for the frame rate is 100. The classifier used for speech emotion recognition is the Multi-Layer Perceptron (MLP) (Bishop, 1995). It has the ability to find a non-linear separation of emotional states 8 and has been widely used in recent applications of speech emotion recognition (Al-Ayadi *et al*., 2011). Figure 1 shows the general structure of MLP that is employed for this work. It consists of three basic layers, namely; input, hidden and output layers. In each layer, it has its own specific functions. The input in MLP is the feature vector extracted from the object to be classified. Since Slaney's MFCC feature extraction method produces 40 features, the input neuron for MLP is 40 well. Based on the preliminary experiments (Kamaruddin *et al*., 2012; Kamaruddin and Wahab, 2013), 2 layer hidden with 10 neurons architecture is implemented with four neurons in the output layer to segregate the four different emotion of anger, happiness, sadness and neutral (Fig. 1). MLP Architecture For normalization, 1000 instances are randomly selected from each emotion. Since there are four speakers for Netherland EmoSpeech dataset, 250 instances are picked from each speaker in order to minimize bias (250 instancesx4 speakers = 1000 instances). Furthermore, 5-fold validation technique is implemented. Such approach is conducted to ensure the classifier produce generalization result rather than memorization (using the same data for training and testing). Training is the process to familiarize the classifier with the distinct emotion characteristics that differentiate between one emotions to another while testing is the process to check the similarity match between the trained classifier and the newly introduced instances. The result are then recorded as accuracy performance. The 80% of the data (3200 instances) are used for training and the remaining 20% are reserved for testing. Such training-testing pairs are repeated until the whole data have been used.

RESULTS AND DISCUSSION

The speech emotion recognition system is developed using the relevant features extracted from MFCC coupled with MLP classifier. For accuracy performance, the experimental results for NTU-American and Netherlands EmoSpeech datasets are presented in Fig. 2. Both datasets managed to yield accuracy between 55 and 70% which is two times higher than chance guessing (25% for 4-classes identification task). Such results signify that the implemented approach is able to recognize emotion with comparative results. From Fig. 2, it is observed that sadness is consistently highly recognized emotion which obtained the highest performance with 69.8 and 64.8% accuracy for NTU-American and Netherland EmoSpeech datasets, respectively. However, happiness recorded the worst performance in NTU-American dataset with 57.4% that differ 12.4% from sadness accuracy in the same dataset. Such trend is not similar to Netherland EmoSpeech performance where neutral reigned the lowest accuracy of 56.3% and the difference of 6.5%. The mean overall performance for NTU-American is 65.5% whereas 59.1% for Netherlands EmoSpeech dataset. Detailed results for different speakers for the Netherlands EmoSpeech dataset are also recorded and presented in Fig. 3. It shows that individual speech emotion classification result is much better compared to cumulative speech emotion classification result. This is because once we combined different speaker's data to the data pool, the complexity for the classifier task is increased resulting to lower accuracy performance. From the result in Fig. 3, speaker 3 emotions are easily identified compared to the other speakers with overall mean performance of 83.9%. Speaker 4 and Speaker 1 follow the ranking with 76.0 and 73.8% overall mean performance respectively. The lowest accuracy is recorded by Speaker 2 with variance value of 16.2% than result Speaker 3. There is no obvious pattern for highest speech emotion classification for individual performance in Netherlands EmoSpeech dataset. Sadness is recorded the highest for Speaker1 (85.9%) and Speaker 3 (93.3%). However, the highest performance for Speaker 2 is happiness (74.5% accuracy) and neutral (87.5% accuracy) for Speaker 4. The same observation is also noted for the lowest speech emotion classification for individual performance. Sadness, albeit the highest accuracy for Speaker 1 and 3 recorded the worst performance for Speaker 2 with 57.6% accuracy. Anger in Speaker 1 and 4 yielded the lowest accuracy with 63.5 and 64.9%, respectively. In summary, based on the results in Fig. 2, once the data of individuals are added to the data pool, the distinction of one individual to another is suppressed giving only the

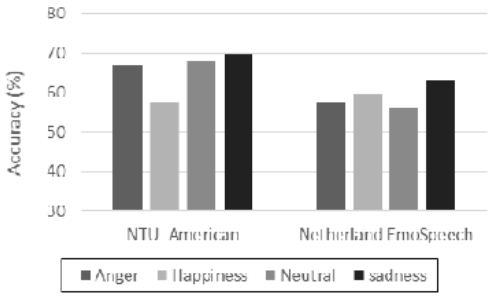

Fig. 2: Speech emotion recognition accuracy for NTU-American and Netherlands EmoSpeech datasets

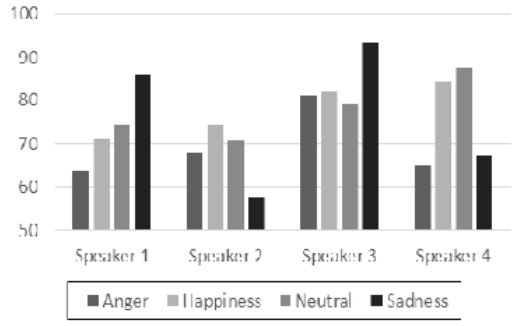

Fig. 3: Speech emotion recognition accuracy for different speakers in Netherlands EmoSpeech dataset

effect of collective data (cultural effect). The accuracy for collective recognition however is lower than individual speech emotion recognition results as presented in Fig. 3. Such result indicates that culture gives influence to the overall accuracy of speech emotion recognition. Moreover, the performance of Netherlands EmoSpeech is comparable to NTU-American with only 6.4% difference. Such5 result indicate that the newly collected dataset of Netherlands EmoSpeech can be used for speech emotion recognition task.

CONCLUSION

Speech is the medium for human communication and it is one of the signal transaction process. However, to complicate matters, culture plays an important role in affecting the way the speech is delivered. Intra-cultural assessment manages to score better accuracy than compared to inter-cultural assessment because of the unique and distinct subtle encoding and decoding information that is shared among the same cultural group. However, universality of emotion allows one to correctly classify emotion although the understanding of the word's semantic is not available. In this study, two

different datasets, NTU-American and Netherlands EmoSpeech are used to represent American and European cultural influence. The heterogeneous assessment is conducted by forty Asian judges who does not have semantic understanding of the words uttered by the inter-cultural group speakers. Such measure is taken to ensure that the inference is done on the acoustic perception only without the interference of semantic knowledge. The MFCC feature extraction method is coupled with MLP classifier to determine four different emotions, namely; anger, happiness, sadness and neutral acting as emotionless state. The experimental results shown that approach has potential for further exploration and analysis for better performance with accuracy performance ranging from 55-70% for cumulative classification accuracy and 55-93% for individual classification accuracy. The disparity between cumulative and individual classification performance is because the complexity for classifier is increased once the different individual data is added to the data pool for training and testing. Moreover, an introduction to the Netherland EmoSpeech dataset provides additional data for speech emotion recognition analysis. More effort should be focused on the use of different feature extraction methods (Kamaruddin and Waab, 2013) and classifiers to yield better accuracy. In addition, cultural effect on speech emotion recognition cannot be ignored to gain optimum performance on speech emotion recognition system. The extension of the analysis can be used for human behavior study such as driving behavior (Khalid *et al*., 2008), customer satisfaction (Kamaruddin and Waab, 2013) and brain signal interaction to emotion (Wahab *et al*., 2010).

ACKNOWLEDGEMENTS

Researchers would like to thank Universiti Teknologi MARA Malaysia (UiTM) and Ministry of Education Malaysia (KPM) for providing financial support through the Research Acculturation Grant Scheme RAGS (600 B RMI / RAGS 5/3 (5/2014)) to conduct the work published in this study.

REFERENCES

Beaupre, M.G. and U. Hess, 2005. Cross-cultural emotion recognition among Canadian ethnic groups. J. Cross-Cult. Psychol., 36: 355-370.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK., ISBN-13: 9780198538646, Pages: 482.

Dewaele, J.M., 2008. The emotional weight of *I love you* in multilinguals' languages. J. Pragmatics, 40: 1753-1780.

Ekman, P., 1971. Universals and Cultural Differences in Facial Expressions of Emotion. In: Nebraska Symposium on Motivation, Cole, J.K. (Ed.). University of Nebraska Press, Lincoln, NE., USA., ISBN-13: 978-0803256194, pp: 207-283.

El Ayadi, M., M.S. Kamel and F. Karray, 2011. Survey on speech emotion recognition: Features, classification schemes and databases. Pattern Recognit., 44: 572-587.

Elfenbein, H.A. and N. Ambady, 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychol. Bull., 128: 203-235.

Kamaruddin, N. and A. Waab, 2013. Measuring customer perceptions index using CMAC speech emotion mapping. Proceedings of the International Conference on Advanced Computer Science Applications and Technologies, December 23-24, 2013, Kuching, Malaysia, pp: 352-357.

Kamaruddin, N. and A. Wahab, 2009a. Features extraction for speech emotion. J. Comput. Methods Sci. Eng., 9: 1-12.

Kamaruddin, N. and A. Wahab, 2009b. CMAC for speech emotion profiling. Proceedings of the 10th Annual Conference of the International Speech Communication Association, September 6-10, 2009, Brighton, UK -.

Kamaruddin, N. and A. Wahab, 2012. Human behavior state profile mapping based on recalibrated speech affective space model. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 28-September 1, 2012, San Diego, CA., USA., pp: 2021-2024.

Kamaruddin, N., A. Wahab and C. Quek, 2012. Cultural dependency analysis for understanding speech emotion. Expert Syst. Applic., 39: 5115-5133.

Matsumoto, D., 2001. Culture and Emotion. In: Handbook of Culture and Psychology, Matsumoto, D. (Ed.). Oxford University Press, New York, USA., pp: 171-194.

Miyamoto, Y., Y. Uchida and P.C. Ellsworth, 2010. Culture and mixed emotions: Co-occurrence of positive and negative emotions in Japan and the United States. Emotion, 10: 404-415.

Scherer, K.R., R. Banse and H.G. Wallbott, 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cult. Psychol., 32: 76-92.

Slaney, M., 1998. Auditory toolbox, version 2. Technical Report No. 1998-010, Interval Research Corporation, USA.

Wahab, A., N. Kamaruddin, L.K. Palaniappan, M. Li and R. Khosrowabadi, 2010. EEG signals for emotion recognition. J. Comput. Methods Sci. Eng., 10: 1-11.