GENDER DEPENDENT WORD-LEVEL EMOTION DETECTION USING GLOBAL SPECTRAL SPEECH FEATURES

HARIS SIDDIQUE

MASTER OF SCIENCE
UNIVERSITI UTARA MALAYSIA
2015

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Malaysia

Abstrak

Dalam kajian ini , ciri-ciri spektrum global dipetik daripada perkataan dan ayat yang dikaji untuk pengecaman pertuturan emosi. MFCC (Mel Kekerapan Cepstral Pekali) telah digunakan sebagai maklumat spektrum untuk tujuan pengecaman. Ciri spektrum global mewakili statistik kasar seperti purata MFCC digunakan. Kajian ini juga mengkaji perkataan di kedudukan yang berbeza (di awal, pertengahan dan akhir) secara berasingan dalam ayat. Ciri pengekstrakan tahap perkataan digunakan untuk menganalisa prestasi pengecaman emosi berdasarkan perkataan di kedudukan yang berbeza. Sempadan perkataan dikenalpasti secara manual. Model berdasarkan jantina atau model bebas jantina juga dikaji untuk menganalisa kesan jantina ke atas prestasi pengecaman emosi. Berlin Emo - DB (Pangkalan Data emosi) telah digunakan sebagai set data ucapan beremosi. Prestasi pengklasifikasi-pengklasifikasi yang berbeza juga dikaji. NN (Rangkaian Neural), KNN (K - Jiran Terdekat) dan LDA (Analisa Diskriminasi Linear) adalah pengklasifikasi yang digunakan. Emosi kemarahan dan neutral juga dikaji. Keputusan menunjukkan bahawa, dengan menggunakan semua 13 pekali MFCC memberikan hasil yang lebih baik daripada pengelasan gabungan lain pekali MFCC untuk emosi yang dinyatakan. Perkataan-perkataan di kedudukan permulaan dan berakhir menandakan posisi emosi lebih baik daripada kandungan emosi di kedudukan pertengahan. Prestasi model berdasarkan jantina adalah lebih baik daripada jantina model bebas jantina. Selain itu, wanita adalah lebih baik daripada lelaki dari segi mempamerkan emosi. Secara amnya, prestasi NN adalah paling teruk daripada KNN dan LDA dari segi klasifikasi emosi marah dan neutral. Prestasi LDA adalah lebih baik daripada KNN sebanyak hampir 15% dengan menggunakan model bebas jantina dan hampir 25% menggunakan model berdasarkan jantina.

Kata Kunci: Koefisyen Kekerapan Mel Cepstral, pengekstrakan ciri, pengecaman ucapan beremosi, korpus simulasi ucapan beremosi, model klasifikasi

Abstract

In this study, global spectral features extracted from word and sentence levels are studied for speech emotion recognition. MFCC (Mel Frequency Cepstral Coefficient) were used as spectral information for recognition purpose. Global spectral features representing gross statistics such as mean of MFCC are used. This study also examine words at different positions (initial, middle and end) separately in a sentence. Word-level feature extraction is used to analyze emotion recognition performance of words at different positions. Word boundaries are manually identified. Gender dependent and independent models are also studied to analyze the gender impact on emotion recognition performance. Berlin's Emo-DB (Emotional Database) was used for emotional speech dataset. Performance of different classifiers also been studied. NN (Neural Network), KNN (K-Nearest Neighbor) and LDA (Linear Discriminant Analysis) are included in the classifiers. Anger and neutral emotions were also studied. Results showed that, using all 13 MFCC coefficients provide better classification results than other combinations of MFCC coefficients for the mentioned emotions. Words at initial and ending positions provide more emotion, specific information than words at middle position. Gender dependent models are more efficient than gender independent models. Moreover, female are more efficient than male model and female exhibit emotions better than the male. General, NN performs the worst compared to KNN and LDA in classifying anger and neutral. LDA performs better than KNN almost 15% for gender independent model and almost 25% for gender dependent.

Keywords: Mel Frequency Cepstral coefficents, Feature extraction, emotional speech recognition, simulated emotional speech corpus, classification models

Table of Contents

List of Tables

List of Figures

CHAPTER ONE
INTRODUCTION

1.1 Background

Speech is the fastest medium of presenting messages in face-to-face communication. On the other hand, emotions are also another medium of communication, for example, a smaller set of gestures can describe the individual's emotional state to others. One debatable topic on it is that a smile or laugh is treated as a signal of happiness in all civilizations. Whereas, crying is treated every bit a sign of sadness or heartbreak, their assessments can change from culture to culture (Lewis et al., 2008).

Most human has emotional activities and we often produce emotions in our free time. Usually we take novels, films, music or other plays as a root of our amusement. These all portray real emotions, but about unreal events, and beside we know that these are fictions or unreal, we make an emotional attachment with them. Reason is the means we used to interact with or understand with everyday real world, we practice the same method with these informants, that's why a particular music makes us happy and other character of music makes us sad (Lewis et al., 2008).

According to Ling He (2010) emotions are psychological and physiological states that take in actual and liberated responses. Emotions comprise person's state of mind and the way a person interacts with others as well as with the environment. Sometimes emotion is related to 'mood'. Normally emotion is short-timed physiological and psychological state that last from a few minutes to a few hours, whereas mood is long-timed emotional state that can last from hours to weeks (Ling He, 2010).

The contents of the thesis is for internal user only

References

Agarwal, A., Jain, A., Prakash, N., &Agrawal, S. S. (2010). Word boundary detection in continuous speech based on suprasegmental features for hindi language. *2010 2nd International Conference on Signal Processing Systems*, V2–591–V2–594. doi:10.1109/ICSPS.2010.5555691

Anagnostopoulos, C.-N.,&Iliou, T. (2010). Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research. In M. Wallace, I. Anagnostopoulos, P. Mylonas& M. Bieliková (ed.), *Semantics in Adaptive and Personalized Services* , Vol. 279 (pp. 127-143) . Springer . ISBN: 978-3-642-11683-4.

Bajpai, A., &Yegnanarayana, B. (2008). Combining evidence from subsegmental and segmental features for audio clip classification. *TENCON 2008 - 2008 IEEE Region 10 Conference*, 1–5. doi:10.1109/TENCON.2008.4766692

Bapineedu, G., Avinash, B., Gangashetty, S. V., &Yegnanarayana, B. (2009). Analysis of lombard speech using excitation source information. In *INTERSPEECH-09*, Brighton, UK, 6–10 September 2009 (pp. 1091–1094).

Bitouk, D., Verma, R.,&Nenkova, A. (2010). Class-level spectral features for emotion recognition.*Speech Communication*, 52, 613-625.

Bozkurt, E., Erzin, E., Erdem, Ç. E.,&Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals.. *INTERSPEECH* (p./pp. 324-327), : ISCA.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F.,& Weiss, B. (2005). A database of German emotional speech.. *INTERSPEECH* (p./pp. 1517-1520), : ISCA.

Chauhan, A., Koolagudi, S. G., Kafley, S., &Rao, K. S. (2010). Emotion recognition using lp residual. In *IEEE TechSym 2010*, West Bengal, India, April 2010. IITKharagpur: IEEE.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*(1-2), 33–60. doi:10.1016/S0167-6393(02)00070-5

Gaurav, M. (2008). Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech.. In A. Das & S. Bangalore (eds.), *SLT* (p./pp. 313-316), : IEEE. ISBN: 978-1-4244-3472-5

Han, Z., Lun, S., & Wang, J. (2012). A Study on Speech Emotion Recognition Based on CCBC and Neural Network. *2012 International Conference on Computer Science and Electronics Engineering*, 144–147. doi:10.1109/ICCSEE.2012.128

Hanjalic, a. (2005). Affective video content representation and modeling. *IEEE Transactions on Multimedia*, *7*(1), 143–154. doi:10.1109/TMM.2004.840618

Hansen, J. H. L., Bou-Ghazale, S. E., Sarikaya, R., &Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database.. In G. Kokkinakis, N. Fakotakis& E. Dermatas (eds.), *EUROSPEECH*, : ISCA.

Hansen, J. H. L.,Sangwan, A., & Kim, W. (2012). Speech Under Stress and Lombard Effect - Impact and Solutions for Forensic Speaker Recognition. Forensic Speaker Recognition. (A. Neustein& H. A. Patil, Eds.). doi:10.1007/978-1-4614-0263-3

Henríquez Rodríguez, P., Alonso Hernández, J. B., FerrerBallester, M. a., Travieso González, C. M., & Orozco-Arroyave, J. R. (2012). Global Selection of Features for Nonlinear Dynamics Characterization of Emotional Speech. *Cognitive Computation*, *5*(4), 517–525. doi:10.1007/s12559-012-9157-0

Henríquez, P., Alonso, J. B., Ferrer, M. A., Travieso, C. M.,& Orozco-Arroyave, J. R. (2011). Application of Nonlinear Dynamics Characterization to Emotional Speech.. In C. M. Travieso-González & J. B. A. Hernández (eds.), *NOLISP* (p./pp. 127-136), : Springer. ISBN: 978-3-642-25019-4

Iida, A., Campbell, N., Higuchi, F., &Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, *40*(1-2), 161–187. doi:10.1016/S0167-6393(02)00081-X

Iliev, A. I., Scordilis, M. S., Papa, J. P., &Falcão, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*, *24*(3), 445–460. doi:10.1016/j.csl.2009.02.005

Iliou, T., &Anagnostopoulos, C.-N. (2009). Comparison of Different Classifiers for Emotion Recognition. *2009 13th Panhellenic Conference on Informatics*, 102–106. doi:10.1109/PCI.2009.7

Iliou, T., &Anagnostopoulos, C.-N. (2009b). Statistical Evaluation of Speech Features for Emotion Recognition. *2009 Fourth International Conference on Digital Telecommunications*, 121–126. doi:10.1109/ICDT.2009.30

Kamaruddin, N., &Wahab, A. (2009). Features extraction for speech emotion. *Journal of Computational Methods in Science and Engineering*,*9*(9), 1–12.

Khanna, P.,& Kumar, M. S. (2011). Application of Vector Quantization in Emotion Recognition from Human Speech.. In S. Dua, S. Sahni& D. P. Goyal (eds.), *ICISTM* (p./pp. 118-125), : Springer. ISBN: 978-3-642-19422-1

Kishore, K. V. K., &Satish, P. K. (2013). Emotion Recognition in Speech Using MFCC and Wavelet Features. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International* , vol., no., pp.842,847

Kodukula, R. SRI. (2009).Significance of excitation source information for speech analysis. Unpublished doctoral dissertation,Dept. of Computer Science and Engineering, IIT Madras, Chennai, India.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, *15*(2), 265–289. doi:10.1007/s10772-012-9139-3

Koolagudi, S. G., & Rao, K. S. (2012b). Emotion recognition from speech: a review. *International Journal of Speech Technology*, *15*(2), 99-117.

Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011). Speech Emotion Recognition Using Segmental Level Prosodic Analysis. *2011 International Conference on Devices and Communications (ICDeCom)*, 1–5. doi:10.1109/ICDECOM.2011.5738536

Koolagudi, S. G., Maity, S., Vuppala, A. K., Chakrabarti, S.,&Rao, K. S. (2009). IITKGP-SESC: Speech Database for Emotion Analysis. In S. Ranka, S. Aluru, R. Buyya, Y.-C. Chung, S. Dua, A. Grama, S. K. S. Gupta, R. Kumar & V. V. Phoha (eds.), *IC3* (p./pp. 485-492), : Springer. ISBN: 978-3-642-03546-3

Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. *International Conference on Signal Processing and Communications (SPCOM)*, 1–5. doi:10.1109/SPCOM.2010.5560541

Koolagudi, S. G., Reddy, R., Yadav, J., & Rao, K. S. (2011b). IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In *Devices and Communications (ICDeCom), 2011 International Conference on* (pp. 1-5). IEEE.

Kuchibhotla, S., Yalamanchili, B. S., Vankayalapati, H. D., & Anne, K. R. (2014, January). Speech Emotion Recognition Using Regularized Discriminant Analysis. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013* (pp. 363-369). Springer International Publishing.

Lee, C. M.,& Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs.. *IEEE Transactions on Speech and Audio Processing*, 13, 293-303.

Lewis, M., Haviland-Jones, J. M., & Barrett, L. F., (2008). *Handbook of Emotions.* Third Edition. The guilford press ISBN: 1609180445

Ling He. (2010). Stress and Emotion Recognition in Natural Speech in the Work and Family Environments. Unpublished doctoral dissertation, School of Electrical and Computer Engineering,RMIT University.

Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *ICASSP*, Honolulu, Hawaii, USA, May 2007 (pp. IV17–IV20). New York: IEEE Press.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580.

Mao, X., Chen, L., & Fu, L. (2009). Multi-level Speech Emotion Recognition Based on HMM and ANN. *2009 WRI World Congress on Computer Science and Information Engineering*, 225–229. doi:10.1109/CSIE.2009.113

Mubarak, O. M., Ambikairajah, E., & Epps, J. (2005). Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. In *8th international symposium on signal processing and its applications*, Sydney, Australia, Aug. 2005.

MY, S. A. (2014). An Improved Feature Extraction Method for Malay Vowel Recognition based on Spectrum Delta.

Neiberg, D., Elenius, K., &Laskowski, K. (2006). Emotion Recognition in Spontaneous Speech Using GMMs Department of Speech , Music and Hearing , KTH , Stockholm , Sweden Classifiers, 809–812.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*(4), 603–623. doi:10.1016/S0167-6393(03)00099-2

Pao, T.-L., Chen, Y.-T., Yeh, J.-H., Cheng, Y.-M.,&Chien, C. S. (2007). Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech.. In A. Paiva, R. Prada & R. W. Picard (eds.), *ACII* (p./pp. 741-742), : Springer. ISBN: 978-3-540-74888-5

Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. (2005). Combining acoustic features for improved emotion recognition in mandarin speech. In *Affective Computing and Intelligent Interaction* (pp. 279-285). Springer Berlin Heidelberg.

Pao, T. L., Liao, W. Y., & Chen, Y. T. (2008). A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition. *Speech Recognition Technologies and Applications*, 550-552.

Rabiner, L. R., Juang, B.-H. (1993). *Fundamentals of speech recognition.*. Prentice Hall. ISBN: 978-0-13-015157-5

Ramakrishnan, S., &Emary, I. M. M. (2011). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, *52*(3), 1467–1478. doi:10.1007/s11235-011-9624-z

Rao, K. S., Koolagudi, S. G., &Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143–160. doi:10.1007/s10772-012-9172-2

Rao, K. S., Prasanna, S. R. M., &Sagar, T. V. (2007b). Emotion recognition using multilevel prosodic information. In *Workshop on image and signal processing (WISP-2007)*, Guwahati, India, Dec. 2007. Guwahati: IIT Guwahati.

Rao, K. S., Prasanna, S. R. M., Yegnanarayana, B., & Member, S. (2007a). Determination of Instants of Significant Excitation in Speech Using Hilbert Envelope and Group Delay Function, *14*(10), 762–765.

Rao, K. S., Reddy, R., Maity, S., &Koolagudi, S. G. (2010) Characterization of emotions using the dynamics of prosodic features. School of Information Technology Indian Institute of Technology Kharagpur.

Rao, K. S., & Koolagudi, S. G. (2012). *Emotion Recognition Using Speech Features*. Springer.

Rojas, R. (1996). *Neural networks: a systematic introduction*. Springer.

Rong, J., Li, G., & Chen, Y.-P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*, *45*(3), 315–328. doi:10.1016/j.ipm.2008.09.003

Scherer, S., Schwenker, F.,& Palm, G. (2008). Emotion Recognition from Speech Using Multi-Classifier Systems and RBF-Ensembles.. In B. Prasad & S. R. M. Prasanna (ed.), *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks* , Vol. 83 (pp. 49-70) . Springer . ISBN: 978-3-540-75397-1.

Schroder, M., &Cowie, R. (2006). Issues in emotion-oriented computing toward a shared understanding. In *Workshop on emotion and computing (HUMAINE)*.

Sethu, V., Ambikairajah, E., & Epps, J. (2013). On the use of speech parameter contours for emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 19. doi:10.1186/1687-4722-2013-19

Shi, Y., & Song, W. (2010). Speech emotion recognition based on data mining technology. *2010 Sixth International Conference on Natural Computation*, (Icnc), 615–619. doi:10.1109/ICNC.2010.5583142

Staroniewicz, P. (2009). Recognition of Emotional State in Polish Speech - Comparison between Human and Automatic Efficiency, 33–40.

Staroniewicz, P. (2011). Automatic Recognition of Emotional State in Polish Speech.. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller & G. Scarpetta (eds.), *COST 2102 Training School* (p./pp. 347-353), : Springer. ISBN: 978-3-642-18183-2

Staroniewicz, P. (2011a). Influence of Speakers' Emotional States on Voice Recognition Scores.. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud& A. Nijholt (eds.), *COST 2102 Conference* (p./pp. 223-228), : Springer. ISBN: 978-3-642-25774-2

Tao, J.,& Kang, Y. (2005). Features Importance Analysis for Emotional Speech Classification.. In J. Tao, T. Tan & R. W. Picard (eds.), *ACII* (p./pp. 449-457), : Springer. ISBN: 3-540-29621-2

Ververidis, D.,&Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods.. *Speech Communication*, 48, 1162-1181.

Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *ICASSP* (pp. I593–I596). New York: IEEE Press.

Vicsi, K.,&Sztahó, D. (2011). Problems of the Automatic Emotion Recognitions in Spontaneous Speech; An Example for the Recognition in a Dispatcher Center.. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller & G. Scarpetta (eds.), *COST 2102 Training School* (p./pp. 331-339), : Springer. ISBN: 978-3-642-18183-2

Wang, J., Han, Z., & Lung, S. (2011, October). Speech emotion recognition system based on genetic algorithm and neural network. *International Conference on Image Analysis and Signal Processing (IASP)*, (pp. 578-582).

Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and Optimal Classification of Speech Emotion Recognition. *2008 Fourth International Conference on Natural Computation*, 407–411. doi:10.1109/ICNC.2008.713

Wang, Y., & Guan, L. (2004). An investigation of speech-based human emotion recognition. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on* (pp. 15-18). IEEE.

Wu, C., Member, S., & Liang, W. (2011). Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels, *2*(1), 10–21.

Wu, C., Yeh, J., & Chuang, Z. (2009b). Emotion perception and recognition from speech. *Affective Information Processing*. Retrieved from http://link.springer.com/chapter/10.1007/978-1-84800-306-4_6

Wu, S., Falk, T.H., & Chan, W. Y., (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. *Digital Signal Processing, 2009 16th International Conference on* , vol., no., pp.1,6, 5-7 July.

Yegnanarayana, B., Member, S., Swamy, R. K., &Murty, K. S. R. (2009). Determining Mixing Parameters From Multispeaker Data Using Speech-Specific Information, *17*(6), 1196–1207.

Yong-Wan, R., Kim, D.-J., Lee, W.-S., & Hong, K.-S. (2009). Novel acoustic features for speech emotion recognition. *Science in China Series E: Technological Sciences*, *52*(7), 1838–1848. doi:10.1007/s11431-009-0204-3

You, M., Chen, C.,Bu, J., Liu, J., Tao, J. (2007). Emotion recognition from speech signals combining PCA and LDA**.** *17*(6), 1196–1207.

Zhang, H. (2012). Emotional Speech Recognition Based on Syllable Distribution Feature Extraction, In Advances in Intelligent and Soft Computing (AISC) Volume 122, 2012, pp 415-420.

Zhao, X., Zhang, S., & Lei, B. (2013). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*. doi:10.1007/s00521-013-1377-z

Zhiyan, H., &Jian, W. (2013). Speech emotion recognition based on wavelet transform and improved HMM. *2013 25th Chinese Control and Decision Conference (CCDC)*, 3156–3159. doi:10.1109/CCDC.2013.6561489

Zhou, Y., Sun, Y., Yang, L., & Yan, Y. (2009). Applying Articulatory Features to Speech Emotion Recognition. *2009 International Conference on Research Challenges in Computer Science*, 73–76. doi:10.1109/ICRCCS.2009.26