Abstract

Speech emotion recognition (SER) system is becoming a very important tool for human-computer interaction. Previous studies in SER have been focused on utterance as one unit. They assumed that the emotional state is fixed during the utterance, although, the emotional state may change during the time even in one utterance. Therefore, using utterance as one unit is not suitable for this purpose especially for long utterances. The ultimate goal of this study is to find a novel emotion unit that can be used to improve SER accuracy. Therefore, different emotion units defined based on voiced segments are investigated. To find the optimal emotion unit, SER system based on support vector machine (SVM) classifier is used to evaluate each unit. The classification rate is used as a metric for the evaluation. To validate the proposed method, the Berlin database of emotional speech EMO-DB is used. The experimental results revealed that emotion unit that contains four voiced segments gives the highest recognition rate for SER. Moreover, the final emotional state of the whole utterance is determined by majority voting of emotional states of its units. It is found that the performance of the proposed method using voiced related emotion unit outperforms the conventional method using utterance as one
References

1. Jiang W., Wang Z., Jin J.S., Han X., Li C. Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network. Sensors (Basel). 2019 Jun 18;19(12):2730.
2. Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., and Salichs, M. A., "A multimodal emotion detection system during human–robot interaction," Sensors, vol. 13, no. 11, pp. 15 549–15 581, 2013.
3. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., and Narayanan, S.S., "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.
4. Vogt, T., André, E., "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," In: Proceedings of International Conference on Multimedia & Expo., Amsterdam, The Netherlands (2005).
5. Vogt, T., André, E., and Wagner, J., "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization," In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2007)
6. Kerkeni, L., Serrestou, Y., Raoof, K., Cléder, C., Mahjoub, M., Mbarki, M. (2019). "Automatic Speech Emotion Recognition Using Machine Learning"
7. Battliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: "How to find trouble in communication," Speech Communication 40, 117–143 (2003)
8. Battliner, A., Seppi, D., Steidl, S., and Schuller, B., "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," Advances in Human-Computer Interaction, 2010. vol. 2010, Article ID 782802, 15 pages.
9. Seppi, D., Battliner, A., Steidl, S., Schuller, B., and Noth, E., "Word Accent and Emotion," In Proc. Speech Prosody 2010, Chicago, IL, 2010.
10. Burkhardt, F., Paeschke, A., Rolifes, M.; Sendlmeier, W., and Weiss, B., "A Database of German Emotional Speech," INTERSPEECH (2005).
11. Vlasenko, B., Philippou-Hubner, D., Prylipko, D., Bock, R., Siegert, I., and Wendemuth, A., "Vowels formants analysis allows straightforward detection of high arousal emotions," in 2011 IEEE International Conference on Multimedia and Expo (ICME), 2011.
12. Ringeval, F., & Chetouani, M. (2008). "A vowel based approach for acted emotion recognition," In INTERSPEECH 2008 9th annual conference of the international speech communication association, Brisbane, Australia, 22–26 September.
13. Deb S., and Dandapat, S., "Emotion classification using segmentation of vowel-like and non-vowel-like regions," IEEE Transactions on Affective Computing, 2017
14. Moattar, M., and Homayounpour, M., "A simple but efficient real-time voice activity detection algorithm," in EUSIPCO. EURASIP, 2009, pp. 2549–2553.
15. Moattar, M., Homayounpour, M., and Kalantari, N., "A New Approach for Robust Real-time Voice Activity Detection Using Spectral Pattern," ICASSP, 2010, pp. 4478-4481.
16. H. Kawahara, and I.M.-katsuse, and A.D. Cheveign, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,"
Speech Communication, vol. 27, pp. 187–207, 1999.
17. El Ayadi, M., Kamel, M. S., Karray, F., “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition 44 (3) (2011) 572–587.
18. Busso, C., Lee, S., and Narayanan, S., “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” IEEE Trans. Audio, Speech Language Process., vol. 17, no. 4, pp. 582–596, May 2009.
19. Lee, C., M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., “Emotion recognition based on phoneme classes,”, in Proc. of ICSL, 2004.
20. Eyben, F., Wollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., and Cowie. R., “On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues,” Journal on Multimodal User Interfaces, 3(1-2):7–19, Mar. 2010.
21. Mansoorizadeh M., Charkari N.M., (2007) “Speech emotion recognition: comparison of speech segmentation approaches,” In: Proceedings of IKT, Mashad, Iran
22. Elbarougy, R. and Akagi, M. “Improving Speech Emotion Dimensions Estimation Using a Three-Layer Model for Human Perception,” Journal of Acoustical Science and Technology, 35, 2, 86-98, March, 2014.
23. Elbarougy, R. and Akagi, M., “Optimizing fuzzy inference systems for improving speech emotion recognition,” Advances in Intelligent Systems and Computing, vol. 533, pp. 85-95, 2017.
24. R. Elbarougy, M. Akagi, “Feature selection method for real-time speech emotion recognition,” in Co-ordination and Standardization of Speech Databases and Assessment Techniques (CO-COSDA), 2017 20th Oriental Chapter of the International Committee for the IEEE, 2017, pp. 1–6.

Index Terms

Computer Science
Signal Processing

Keywords

Acoustic features extraction, discriminative features, speech emotion recognition, voiced segments, unvoiced segments.