ABSTRACT OF DISSERTATION FOR FORMAL HEARING

Title: Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information
Intended Degree: Advanced Science and Technology
Name of Laboratory: Akagi-lab
Student Number: s182001

Research Content

Humans perceive emotion in multimodal ways. Speech is one of the sensory modalities in which emotions can be perceived. Within speech, humans communicate emotion through acoustic and linguistic information. In automatic emotion recognition by computers, known as affective computing, there is a shift from unimodal acoustic analysis to multimodal information fusion. As in human speech emotion perception, computers should be able to perform speech emotion recognition (SER) from bimodal acoustic-linguistic information fusion.

This research aims to investigate the necessity to fuse acoustic with linguistic information for recognizing dimensional emotions. To achieve this goal, three sub-goals were addressed: SER by using acoustic features only, fusing acoustic and linguistic information at the feature level, and fusing acoustic and linguistic information at the decision level.

The first strategy aims at maximizing the potency of recognizing dimensional SER by merely using acoustic information through investigating the region of analysis and the effect of silent pause regions. This study generalizes the effectiveness of means and standard deviations to represent acoustic features and the prediction of the importance of silent pause regions for dimensional SER. In addition, the aggregation of acoustic feature models valence and arousal prediction better than the majority voting method. Although several approaches have been carried out, acoustic-based dimensional SER still has some limitations. The major drawback is the low performance of valence's prediction score.

The second and third strategies aim at improving the valence prediction, investigating the necessity of bimodal information fusion, and evaluating the fusion frameworks for fusing acoustic and linguistic information. Two fusion methods for acoustic-linguistic information fusion are studied namely early-fusion approach and late-fusion approach. At the feature level (FL) or early-fusion approach, two fusion methods are evaluated --- feature concatenation and network concatenation. The FL methods showed significant performance improvements over unimodal dimensional SER. At the decision level (DL) or late-fusion approach, acoustic and linguistic information are trained independently, and the results are fused by support vector machine (SVM) to make the final predictions. Although this proposal is more complex than the previous FL fusion, the results showed improvements over the previous DL approach. These studies revealed the necessity to fuse acoustic with linguistic features for dimensional SER.

Table 1 shows the excerpt of results obtained in this doctoral study. The first row shows performance of SER by using acoustic features only. It can be seen that, among other problems, the major disadvantages of acoustic only SER is the low performance of valence prediction. Adding linguistic information at feature level improved the prediction. Moreover, the late fusion method obtained the highest performance, in which this mechanism is also similar on how humans perceive multimodal information. As a comparison, the last row shows result from other research that not only examined acoustic features but also fuse them with age and gender information.

Table 1: Results from different SER methods; the last row is for comparison

Methods	Valence	Arousal	Dominance
Unimodal acoustic	0.298	0.641	0.460
Early fusion	0.446	0.585	0.508
Late fusion	0.596	0.601	0.499
Aocustic+Age+Gender	*0.715*	*0.392*	*0.539*

This dissertation demonstrates the necessity of fusing acoustic with linguistic information for dimensional speech emotion recognition (SER). Aside from this main goal, three sub-goals were transformed into three strategies to investigate the potential solutions of the following problem in dimensional SER:

1. region of analysis for feature extractions: high-level statistics (HSFs) using mean and standard deviation consistently show more meaningful representations for emotion in speech than low-level descriptors (LLD);
2. effect of silent pause regions: silence regions are predicted to contribute in dimensional speech emotion recognition; either removing silence or using silence feature as an additional feature slightly improves the performance score of the baseline whole speech regions;
3. low valence prediction score on dimensional SER: fusing linguistic information with acoustic information could improve the performance of valence prediction;
4. the necessity of fusing acoustic with linguistic information: consistent and significant performance improvements by fusing both acoustic and linguistic information shows the necessity of fusing both pieces of information;
5. framework for fusing acoustic with linguistic information: the late-fusion (decision-level) approach obtained slightly better performance than an early-fusion (feature-level) approach.

Research Purpose

The originality of this doctoral study is the investigation of fusion of acoustic and linguistic information for dimensional speech emotion recognition. Although this approach of acoustic and linguistic fusion is not new, there are several differences among this research and the previous research: (1) most research only focused on categorical emotion, (2) the types of acoustic and linguistic information are different, (2) the investigation of fusion methods is only limited. These differences can be seen as differences in experiment point of view.

At the philosophical level, this research also differs from the previous studies on speech emotion recognition. In these studies, the belief for speech is about how it is said rather than what is said. This study combines both “how” and “what” information. Acoustic features contain information on how it is being said. Linguistic features contain information on what is being said. Fusing both pieces of information, which are extracted from a speech, will improve the clarity of the message, including the expressed emotion. The perception of emotion will also improve by fusing this bimodal information. The process of fusing acoustic and linguistic information for obtaining degree of emotions is also in line with the concept of data-information-knowlege (DIK) which is widely adopted in information science society.

This research contributes to the following area of speech emotion recognition by introducing novelties and strengthening the previous findings.

1. Acoustic feature extraction

   The author evaluated categorical speech emotion recognition from silence removed speech region. The result suggests that extracting acoustic features from the silence-removed region is better than from the whole speech region. In other publication, the author utilized silence as an additional feature to statistical functions. The results achieved a better score than baseline raw speech. These results suggest that either removing silence or utilizing silence as additional features will lead to a better performance than using the whole acoustic signal. The author also confirms and generalizes the effectiveness of mean and standard deviations of low-level acoustic features for SER. The author also showed that acoustic feature aggregation leads to better performances than output aggregation. Finally, the author found that the acoustic features that perform better in SER will also perform better in song emotion recognition.

2. Information fusion

   The author proposed emotion recognition by fusing acoustic and linguistic information at different levels: feature fusion, networks fusion, and decision fusion. The results significantly improved unimodal dimensional emotion recognition from either acoustic or linguistic information. Furthermore, the author discussed the improvement of valence prediction by linguistic information. The evaluated fusion methods are expanded not only for acoustic and linguistic fusion but also for acoustic and visual information fusion.

3. Classification methods

   Modern classification methods utilized deep learning models. However, the conventional method, such as support vector machine (SVM) and multi-layer perceptron (MLP), are still used in many fields. The author showed that traditional MLP with deeper layers and proper configurations performed better than modern deep learning architecture. For the SER task with deep learning, the author confirmed the need for bigger data size to be fed to deep learning models. The choice of the loss function is a matter in machine/deep learning. The author proposed correlation- based loss function to improve the performance of dimensional SER. The author also evaluated multitask learning (MTL) for predicting valence, arousal, and dominance degrees simultaneously based on this loss function. Furthermore, the author found that recurrent-based neural networks (RNN) are effective for the SER task. More improvements were obtained when this RNN model is combined with the attention model.

This research attained an framework for more accurate dimensional speech emotion recognition by fusing acoustic and linguistic information. Although the results showed significant improvement from the unimodal acoustic analysis and the previous fusion methods, there is a lack for practical implementation. The multi-stage process that gained the best result is expensive to be implemented in real life scenario. The future research should tackle the limitation of this research to accelerate its implementation in real and continuous dimensional speech emotion recognition.

Research Accomplishment (r: reviewed, u: unreviewed)

1. B. T. Atmaja and M. Akagi, “Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning,” APSIPA Trans. Signal Inf. Process., Vol. 9, No. May, p. e17, May 2020. (**r**eviewed)
2. R. Elbarougy, B.T. Atmaja and M. Akagi, “Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM,” Journal of Signal Processing, Vol. 24, No. 6, November 2020. (r)
3. B.T. Atmaja and M. Akagi, “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM,” Speech Communication, vol 126, February 2021, pp 9-21. (r)
4. B.T. Atmaja, K. Shirai, and M. Akagi, “Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text,” International Seminar on Science and Technology, Surabaya, 2019. (r)
5. B. T. Atmaja, K. Shirai, and M. Akagi, “Speech Emotion Recognition Using Speech Feature and Word Embedding,” in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 519–523. (r)
6. B. T. Atmaja and M. Akagi, “Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model,” in 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 40–44. (r)
7. B. T. Atmaja and M. Akagi, “Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4482–4486. (r)
8. B. T. Atmaja and M. Akagi, “The Effect of Silence Feature in Dimensional Speech Emotion Recognition,” in 10th International Conference on Speech Prosody 2020, May, pp. 26–30. (r)
9. B.T. Atmaja and M. Akagi, “Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information, ” in 2020 23nd O-COCOSDA, pp. 166-171. (r)
10. B.T. Atmaja and M. Akagi, “On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers”, TENCON 2020, Osaka, Japan, 2020, pp. 968-972. (r)
11. B.T. Atmaja, Y. Hamada and M. Akagi, “Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion” TENCON 2020, Osaka, Japan, 2020, pp. 1081- 1085. (r)
12. B.T. Atmaja and M. Akagi, “Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition,” in 2020 APSIPA ASC, Auckland, New Zealand, 2020. pp. 325-331. (r)
13. B.T. Atmaja and M. Akagi, “Evaluation of Error and Correlation-based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition,” International Conference on Acoustic and Vibration, Bali, Indonesia, 2020. (r)
14. R. Elbarougy, B.T. Atmaja, and M. Akagi, “Continuous Tracking of Emotional State from Speech Based on Emotion Unit,” In Proceeding ASJ Autumn Meeting, 2018. (u)
15. B.T. Atmaja, R. Elbarougy, and M. Akagi, “RNN-based dimensional speech emotion recognition,” In Proc. ASJ Autumn Meeting, 2019, pp. 743–744. (u)
16. B.T. Atmaja and M. Akagi, “Dimensional Speech Emotion Recognition from Acoustic and Text Features Using Multitask Learning,” In Proc. ASJ Spring Meeting, 2020, pp. 1003–1004. (u)