学位論文の骨子（情報科学）　　年　月　日

Dissertation Outline (Information Science)

氏名 Name	Bagus Tris ATMAJA	学生番号 Student Number	S1820002

主指導教員 Supervisor	Masato AKAGI	印 Seal	副指導教員 Second Supervisor	Masashi UNOKI	印 Seal
□ 副テーマ指導教員 Advisor for Minor Research Project □ インターンシップ指導教員 Advisor for Internship				Kiyoaki SHIRAI	印 Seal

＜博士論文題目（仮）＞ (Tentative) Title of Doctoral Dissertation

Dimensional Speech Emotion Recognition by Combining Acoustic and Linguistic Information

＜研究の目的と効果＞ Research Aim and Impact

Humans perceive information in multimodal ways. Among many modalities, speech is an important modality to perceive emotion. Within speech, not only acoustic information can be extracted but also linguistic information. This linguistic information is commonly extracted via speech-to-text technology. While the conventional paradigm in speech emotion recognition (SER) is performed by using acoustic information only, a new paradigm involves multimodal processing to achieve better performances and mimic human multimodal processing. Motivated by this multimodal human perception, this research aims to propose methods for dimensional SER by combining acoustic and linguistic information. The problem thus exists on how to fuse both information. The following strategies are studied to solve this problem: (1) SER by using acoustic features only, (2) combining acoustic and linguistic information at the feature level, and (3) combining acoustic and linguistic information at the decision level. Their descriptions are given below.

The first strategy aims to maximize the potency of recognizing dimensional emotion from acoustic information only. In this study, several acoustic features sets have been evaluated on both low-level and high-level features. Although high statistical functions might be limited in feature size, this kind of feature might be more informative than a local feature since it represents information within an utterance (by mean values) and captures the dynamic between frames (by standard deviation). This study generalizes the previous finding on the effectiveness of means and standard deviations from a specific feature set to other acoustic feature sets. Another way to maximize acoustic-based SER is by investigating the contribution of silent pause for dimensional emotions. Additionally, since a dataset presented audio file in chunks and emotion labels in utterances, an evaluation of aggregation methods for story-based emotion prediction from chunk-based speech data is also investigated. Showing improvements from the state-of-the-art results, the lower scores of valence than other dimensions are still observed in most experiments.

A method to improve acoustic-based SER, particularly due to the low score of valence prediction, is by fusing acoustic and linguistic information. Linguistic information has been reported more predictive than acoustic information in predicting valence. Two fusing methods for acoustic-linguistic information fusion are studied: early-fusion approach and late-fusion approach. First, the early-fusion approach is adopted for the fusion of acoustic and linguistic information at the feature level (FL). In this FL fusion, two fusion methods are evaluated -- feature concatenation and network concatenation. This study also proposes the use of multitask learning (MTL) to predict valence, arousal, and dominance simultaneously by evaluating the number of parameters in the MTL approach. The early-fusion methods show significant performance improvement over unimodal dimensional SER. Second, the late-fusion approach is proposed by fusing predictions from acoustic and linguistic information at decision level (DL). In this DL fusion, acoustic and linguistic information are trained independently, and the results are fused by SVM to make the final predictions. Although this proposal is more complex than the previous FL fusion, the results show improvements. This second fusion is also feasible for future implementation. Currently, ASR produces text from speech accurately. While acoustic features used to train ASR can be used to train SER simultaneously, the transcription from ASR can generate lexical features which result in linguistic-based emotion recognition. This text's prediction can be fused with acoustic's prediction to improve the SER performance.

This research links the current problem in dimensional SER with its potential solution. In dimensional SER, several strategies have been attempted; however, valence's performance is lower than arousal and dominance due to the lack of valence information in acoustic features. On the other hand, sentiment analysis used linguistic information to predict the polarity of sentiment, which is similar to valence. The combination of acoustic and linguistic information solves this problem. Humans also perceive emotion from multimodal information, in which the computation model attempts to mimic. The FL approach improves the performance of unimodal SER significantly, but the DL approach improves the FL approach's performance slightly. The results devote insights for future strategy in implementing SER, whether to use acoustic-only features (less complex, less accurate), an early-fusion method (more complex, more accurate), or a late-fusion method (most complex, most accurate).

＜研究の概要＞ Research Outline

The study of speech emotion recognition (SER) needs advancement to move from the theoretical side to the application side. Twenty years ago, SER was only a hypothetical technology. Nowadays, SER is already implemented from telephone applications to automotive safety. However, the application of SER still lacks many issues. This doctoral study aims to propose methods to tackle these issues.

The main issue of this study is whether it suffices to use acoustic features for modeling emotions, or it is necessary to combine them with linguistic information. Since linguistic features can be extracted from speech via ASR, it is reasonable to use this linguistic feature for future applications without a need to use other modalities. However, the investigation of SER by using acoustic information only is necessary to obtain the advantages and disadvantages of the current SER system. Among many issues, the following problems of SER were being investigated: (1) low score of valence prediction in dimensional SER, (2) region of analysis for feature extraction, (3) effect of silence on the performance of dimensional SER, and (4) how to fuse acoustic and linguistic information for multimodal fusion.

Prior to discussing the main issues is a literature study of the research theme. This study summarizes current trends in speech emotion recognition, particularly on dimensional SER and research that used both acoustic and linguistic features as information to recognize emotion within the speech. Although this study is a literature review, a deep review is being conducted; hence this part of the study contributes to the speech emotion recognition community by presenting a summary of several approaches, results, and the remaining problems in the SER area. Apart from information science, emotion from a psychological perspective will also be reviewed, which includes several models of emotion theories. Datasets commonly used in speech emotion research are summarized with more emphasis on datasets used in this research. Finally, the most challenging problems in SER are addressed, including their research direction and potential solutions. This literature study opens insights for the next research parts in this doctoral study.

The second study evaluates and proposes several methods to maximize acoustic-only dimensional speech emotion recognition. Based on previous literature study, it is known that SER by using acoustic features only currently suffers from several issues. Among many issues, two important issues are being tackled in this study. First is the region of analysis used for acoustic feature extraction, whether frame-based processing (local features) or utterance-based processing (global features). Second is the contribution of silent pause features. Several acoustic features are examined to evaluate the effectiveness of the region of analysis. In the first issue, it is found the effectiveness of two high-level statistical functions, i.e., mean and standard deviation, in two different feature sets, which is previously effective on GeMAPS feature sets (Schmitt, 2018). On the second issue, it is found weak to moderate correlation between silent pause feature with dimensional emotions using the evaluated calculation. Furthermore, this study evaluates aggregation methods for chunk-based features to represent story-based features and develop frame-based Mel-filterbank features for dimensional SER. The evidence of this study suggests that acoustic information is necessary for predicting dimensional emotions but might be insufficient for accurate prediction.

The third and fourth studies attempt to cope with low valence prediction by fusing acoustic and linguistic information. The third study fuses acoustic and linguistic information at the feature level. Two information fusions of acoustic and linguistic at the early-fusion approach are evaluated: feature concatenation and network concatenation. At feature concatenation, both acoustic and linguistic features are concatenated and fed into the same classifiers (e.g., SVM, DNN). At network concatenation, both features are fed into different networks. Different datasets are used for these approaches, as well as acoustic features sets (GeMAPS, pyAudioAnalysis, LibROSA), linguistic features (word embedding, word2vec, GloVe, FastText, BeRT), and classifiers (LSTM, CNN, MLP, SVM). This third study also introduces multitask learning (MTL) to concurrently predict valence, arousal, and dominance with different numbers of parameters. This study shows proof of the concept that information combination improves recognizing of emotion significantly. While it is known that linguistic information is more predictive than acoustic for valence and the opposite for arousal, it is found that acoustic information is also more predictive for dominance prediction. The combination of bimodal information with proper configurations improves not only valence prediction but also the predictions of all three-dimensional emotion attributes.

The fourth study fuses acoustic and linguistic information at the decision level. Although the purpose of this study is similar to the third study, the motivation is different. The is work is by how humans perceive multimodal information, which is reported in psychological research. In psychological research, it is believed that humans' multimodal information is processed in the late-fusion approach rather than in the early-fusion approach. For instance, it is argued that semantic and vocal are processed independently via a verbal channel and a vocal channel (Berckmoes, 2004). Although both channels' integration remains unclear, there is evidence that acoustic and linguistic information are processed separately in specific brain areas. This knowledge from psychological research is used to build a computer model to recognize human emotion, particularly for predicting valence, arousal, and dominance, by separating acoustic and linguistic information into different networks to represents vocal and verbal channels. Instead of manual adjustment, the integration of both information is turned to support vector regression (SVR). Given the data partition of testing, development, and test, acoustic and linguistic networks used a training set to predict the development set. The prediction of the development set, which is smaller in size than the training set, is input to the SVR. This two-state processing improves the performance of the FL approach marginally.

＜研究の概要（つづき）＞ (continued from previous page)

This research aims to cope with the drawbacks of dimensional SER by combining acoustic and linguistic information. Several studies have been conducted by proposing several strategies. The first study maximizes the potency of acoustic information for predicting valence, arousal, and dominance. The investigation of high-level features, pause silent features, and acoustic feature aggregation achieves comparable performance to state-of-the-art results. However, the low score of valence is challenging to improve by using acoustic information only. Combining acoustic with linguistic information at feature level improves valence, arousal, and dominance prediction. Moreover, mimicking how humans perceive multimodal information by integrating acoustic and linguistic information at decision level improves dimensional SER performance over the feature-fusion approach. These gradual improvements show that the research direction is being carried out in the right way.

In sum, several issues in dimensional speech emotion recognition are being studied. Several methods have been developed, and the improved version of these methods is being crafted. The current results show proof of concept used in this research, i.e., the combination of acoustic and linguistic information for dimensional emotion prediction works well. This research achieves its purpose to answer the main issue about the necessity of combining linguistic information to advantage dimensional SER. Based on the complexities and performances, three approaches can be chosen whether to use acoustic information only or a combination of acoustic and linguistic information. The results show the potential advantage of the proposed methods to advance speech-based emotion recognition technology, particularly on dimensional SER.

＜論文の構成案＞ Dissertation Structure

1. Introduction
    1. Background
    2. Motivation
    3. Research Concept
    4. Research Issues
    5. Organization of Dissertation
2. Literature Review
    1. Psychology of Emotion
    2. Emotion Model
    3. Datasets
    4. Feature Sets
    5. Classifiers
    6. Summary
3. SER by Using Acoustic Information
    1. Introduction
    2. SER using low-level features
    3. SER using high-level features
    4. Feature aggregation methods
    5. Using silent pause as a feature
    6. Summary
4. Early Fusion of Acoustic and Linguistic Information
    1. Introduction
    2. Early fusion by features concatenation
    3. Early fusion by network concatenation
    4. Comparing ASR output with manual transcription
    5. Summary
5. Late Fusion of Acoustic and Linguistic Information
    1. Introduction
    2. Two-stage dimensional SER
        1. LSTM network for unimodal prediction
        2. SVM for modality fusion
    3. Psychologically inspired dimensional SER
    4. Benchmarking results
    5. Summary
6. Conclusions
    1. Summary
    2. Future work

＜研究業績＞ Publication List

Domestic Conferences (unreviewed):

1. Reda Elbarougy, B.T. Atmaja, Masato Akagi, “Continuous Tracking of Emotional State from Speech Based on Emotion Unit”, ASJ Autumn Meeting, Oita, 2018.
2. Atmaja, B.T., Arifianto, D., Akhmad, F., Akagi, M., 2019. “Speech recognition on Indonesian language by using time delay neural network.” ASJ Spring Meeting, Tokyo, pp 1291–1294.
3. Atmaja, B.T., Elbarougy, R., Akagi, M., 2019. “RNN-based Dimensional Speech Emotion Recognition,” in: ASJ Autumn Meeting. Shiga, pp. 743–744.
4. Bagus Tris Atmaja and Masato Akagi, “Dimensional Speech Emotion Recognition from Speech and Text Features using MTL,” in: ASJ Spring Meeting 2020, Saitama, pp. 1003-1004.

International Conferences (reviewed):

1. Atmaja, Bagus Tris, Kiyoaki Shirai, and Masato Akagi, “Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text”, *In 2019 International Seminar on Science and Technology (ISST)*, Surabaya, 2019.
2. Atmaja, Bagus Tris, and Masato Akagi. "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model." In *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pp. 40-44. IEEE, 2019.
3. Atmaja, Bagus Tris, Kiyoaki Shirai, and Masato 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)*, pp. 519-523. IEEE, 2019.
4. Atmaja, Bagus Tris, and Masato 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)*, pp. 4482-4486. IEEE, 2020.
5. Atmaja, B.T., Akagi, M. “The Effect of Silence Feature in Dimensional Speech Emotion Recognition.” Proc. 10th International Conference on Speech Prosody 2020, 26-30, DOI: 10.21437/SpeechProsody.2020-6.

Journals (reviewed):

1. Atmaja, B.T., Akagi, M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. APSIPA Transaction on Signal and Information Processing, vol. 9, e17. DOI: https://doi.org/10.1017/ATSIP.2020.14
2. Reda Elbarougy, Bagus Tris Atmaja, Masato Akagi (2020). “Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM”, Journal of Signal Processing (accepted/to appear).

＜現在の単位修得状況＞ Courses I have obtained credits

	発展科目 Intermediate	先端科目 Advanced	その他 Others	合計 Total	必修 B 科目(S50x) Required Courses
科目数 Number of courses	6	3	3	12	□ S503 □ S501 / S502
単位数 Number of credits	11	6	6	23	