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 process information in multimodal ways. Among many modalities, speech is an important modality in which emotion can be perceived. Within speech, not only acoustic information can be extracted but also linguistic information (via automatic speech recognition, ASR). This research aims to propose methods for combining acoustic and linguistic information for dimensional speech emotion recognition (SER). To achieve these aims, the followings are studied: (1) SER 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. A brief description of each part is given below.

Ahead from combining acoustic and linguistic information for dimensional SER, dimensional emotion recognition with acoustic only information is examined. This study 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. This study contributes to the generalization of high-level statistical functions, i.e., mean and standard deviation, which is previously observed on a specific feature set. The use of these high-level features improves the performance of SER over low-level descriptors. This study also contributes to SER by evaluating the correlation between silent pause features and degree of emotion dimensions and proposing aggregation methods for story-based emotion prediction from chunk-based speech data. One of the drawbacks of acoustic-only SER is the low score of valence compared to other dimensions, which is common in other acoustic-only SER systems.

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 feature-level fusion, two methods to combine acoustic-linguistic information are evaluated, feature concatenation and network concatenation. This study proposes a multitask learning to predict valence, arousal, and dominance simultaneously from bimodal features and evaluates bimodal feature concatenation from story-based linguistic information and acoustic feature aggregation done in the previous study. Using different acoustic features, lexical features, and classifiers, the early-fusion methods show significant performance improvement over single-modality dimensional SER.

Motivated by psychological research results, two-stage processing for dimensional SER via deep learning network (DNN) and support vector machine (SVM) are proposed for fusing acoustic-linguistic information at late-fusion approach. In this study, 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 feature-level fusion, the results show improvement over FL fusion and feasible implementation for future speech technology. 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-based emotion prediction can be fused with acoustic-based emotion prediction to improve SER performance.

This research linked the current problem in dimensional SER with its potential solutions. In dimensional SER, 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. This research achieves its purpose by showing improvements in the performance of each study gradually. 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 application to automotive safety (Nass, 2005). 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 if it is necessary to combine them with linguistic (El-Ayadi, 2011). Since linguistic features can be extracted from speech via ASR, it is reasonable to use this linguistic feature for real future applications without a need to use other modalities. SER using acoustic features only suffers from several issues. Among other issues, two important issues are being tackled in this study. First is the low score of the valence prediction in dimensional SER. Second is the region of analysis used for acoustic feature extraction, whether frame-based processing (local features) or utterance-based processing (global features). In combining acoustic and linguistic information, how to fuse both information is the most important issue, among others.

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

The second part is the first contribution to this research. This research part evaluates and proposes several methods to maximize acoustic-only dimensional speech emotion recognition. Since the target application of this study is speech-based technology, the natural way to investigate the need for such improvements is to study the basic workflow proposed by other researchers. Speech, particularly acoustic and prosody, is known to be related to human emotion. Several acoustic features are examined to evaluate its effectiveness in predicting dimensional emotion within the speech. Next, the region of analysis used for feature extraction is evaluated, i.e., short region and fixed longer region (statistical functions). This part of the research includes the generalization of the effectiveness of two high-level statistical functions, i.e., mean and standard deviation from GeMAPS feature sets (Schmitt, 2018) to several acoustic feature sets. Another important issue tackled by this part of the study is the effect of post-filtering on the speech signal, i.e., removing silence vs. using silence as an acoustic feature. Other evaluations include aggregation methods for chunk-based features to represents story-based features and development of frame-based Mel-filterbank features for dimensional SER, which is being developed. This study suggests that acoustic information is necessary for predicting dimensional emotions but may be insufficient for accurate prediction.

The third part of this research attempts to provide detail of fusing acoustic and linguistic information at the feature level. There is three motivation for conducting this part of the study: the low performance of valence in most DSER, human multimodal perception, and simplicity of early fusion methods. While the first of two motivations are clear, the third motivation, i.e., the proposed methods, is briefly explained here. Two information fusions of acoustic and linguistic at decision level 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) while at network concatenation, both features are fed into different networks (e.g., LSTM, CNN, MLP). Different datasets are used for these approaches, as well as features sets and classifiers. This part of the research also introduces multitask learning (MTL) to concurrently predict valence, arousal, and dominance along with its variants: MTL without parameters, MTL with two parameters, and MTL with three parameters. This study shows proof of concept that information combination improves recognizing of emotion. 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 the recognition of all three dimensional emotion attributes.

The fourth part of this study is the second evaluated method to fuse acoustic and linguistic information, i.e., decision-level fusion. The motivation of the work is the drawbacks of the previous early-fusion method and how humans perceive multimodal information. Given the results from the previous method and motivated by the psychological result on speech emotion perception by acoustic and semantic information, a method was proposed to evaluated dimensional speech emotion recognition by using late fusion or decision-level approach. Psychological and physiological research showed the relation between speech and emotion and how they are processed in our brains. In psychology, research on speech emotion is less developed than in speech processing technology. However, the evidence from psychophysical experiments is strong enough to model human speech emotion speech perception. For instance, it is believed that semantic and vocal are processed independently via a verbal channel and a vocal channel (Berckmoes, 2004). This knowledge from psychological research is used to build a computer model to recognize human emotion, particularly for predicting valence, arousal, and dominance.

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

Motivated by this psychological evidence, a late-fusion dimensional SER from acoustic and linguistic information is proposed. The proposed late-fusion approach is two-stage processing. First, each feature set is trained independently to predict valence, arousal, and dominance by using a long-short term memory network (LSTM). This model represents vocal and verbal channels in the psychological approach. Second, a support vector machine (SVM) is utilized to decide the final prediction of valence, arousal, and dominance from the previous results using LSTM, from both acoustic and linguistic/text networks. The difference in this approach from the previous early-fusion approach will be discussed in the dissertation, as well as its results. In brief, an improvement of using a two-stage late fusion approach is observed over the early-fusion approach.

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 not only contributes to acoustic-linguistic dimensional emotion prediction but also in acoustic-only emotion recognition. In acoustic-only SER, a generalization of high-level features is presented, as well as the contribution of silent pause features and aggregation of acoustic features. In combining acoustic and linguistic information, two different methods are proposed and evaluated. 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
1. Literature Review
    1. Psychology of Emotion
    2. Emotion Model
    3. Datasets
    4. Feature Sets
    5. Classifiers
    6. Summary
2. SER using acoustic features
    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
3. 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
4. Late Fusion of Acoustic and Linguistic Information
    1. Introduction
    2. Two-stage dimensional SER
        a) LSTM network for unimodal prediction
        b) SVM for modality fusion
    3. Psychologically inspired dimensional SER
    4. Benchmarking results
    5. Summary
5. 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	