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 Information2020 6th International Conference on Control, Automation and Robotics (ICCAR)

＜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. In the 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).

In dimensional speech emotion recognition, first, acoustic-only dimensional emotion recognition is studied. This study aims to maximize the potency of dimensional speech emotion recognition from acoustic information only. The contribution of this part of the study is a generalization of high-level statistical functions to some acoustic features sets, which improve performance of SER over low-level descriptors; the correlation between silent pause features and emotion dimensions; and aggregation methods for story-based emotion prediction from chunk-based speech data.

The second part of this research evaluates the fusion of acoustic and linguistic information at the feature level. In this feature-level fusion, two methods to combine acoustic-linguistic information are evaluated, feature concatenation and network concatenation. The contribution of this research part is the proposal of using multitask learning to predict valence, arousal, and dominance simultaneously from bimodal networks and bimodal feature concatenation from story-based linguistic information and acoustic feature aggregation done in the previous part.

The third part proposes two-stage processing for dimensional SER via deep learning network (DNN) and support vector machine (SVM) for fusing acoustic-linguistic information. In this part, 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, as shown in this research.

This research linked the current problem in dimensional SER with its potential solution. In dimensional SER with valence, arousal, and dominance model, the performance of valence 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, which similar to the computer model in several parts, particularly on deciding final emotion recognition via decision-level fusion.

＜研究の概要＞ Research Outline

I am planning to write my dissertation in six chapters. These chapters are Introduction, Literature Review, Speech Emotion Recognition Using Acoustic Features, Combining Acoustic and Linguistic Information at Feature Level, Combining Acoustic and Linguistic Information at Decision Level, and Conclusions. The following is a brief description of those chapters.

In the introduction, I will write the motivation of my research, including the background, the current situation in speech emotion research, and the remaining problems. The concept of tackling these problems is introduced here. Contributions of the research are also presented along with published papers. To guide the readers to easily follow the flow of the writing of the dissertation, an organization of the dissertation will be presented to close this first chapter.

Chapter 2 of the planned dissertation is a review chapter. This chapter will summarize 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 review chapter, I plan to write deeply; hence this chapter 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 will be summarized, including the datasets used in this research. Finally, the most challenging problems in SER are addressed at the end of this chapter, including its potential solution and direction.

Chapter 3-5 is the main contribution to this research. Chapter 3 evaluates and proposes several methods to maximize acoustic-only dimensional speech emotion recognition. This chapter will be based on published papers on this theme. First, related research ideas and their results will be presented. Then, the addressed problem is introduced with its proposed solution. The detailed contribution of this chapter is explained, including a generalization of the effectiveness of high-level acoustic features from several feature sets, evaluation of silent pause features for dimensional speech emotion recognition, and aggregation methods for chunk-based features to represents story-based features. The end of this chapter will summarize works done in this theme, its contribution, and the remaining problems.

Chapter 4 attempts to provide detail of fusing acoustic and linguistic information at the feature level. First, this chapter will briefly review what the remaining problems are. Several potential solutions are introduced, and the proposed method is reviewed in comparison with existing methods. There is three motivation of writing this chapter: 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 chapter 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. The differences among these approaches are explained along with their results. Finally, a summary will end this chapter with some conclusions, contribution, and remaining problems.

In chapter 5, I will write the second evaluated method to fuse acoustic and linguistic information, i.e., decision-level fusion. First, the motivation of the work is introduced, such as the drawbacks of the previous 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. As an introduction, this chapter will show results from psychological research in which the relation between speech and emotion is studied. 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. This knowledge from psychological research can be 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, we presented late-fusion dimensional SER from acoustic and linguistic information in chapter 5. 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, as well as its results. Finally, a summary will close this chapter with some conclusions, contributions, and remaining problems on using the late-fusion approach with two-stage processing for dimensional speech emotion recognition.

The final chapter is a summary. It summarizes all works from chapter 2 to chapter 5. Chapter 5 will present a justification for work done in general. The relation between chapter and its problem is reviewed here. Since science itself is a prediction about nature/universe, including predicting human emotion from speech, the final remarks of this chapter will present research directions to continue research on predicting speech emotion from acoustic and linguistic information.

＜論文の構成案＞ Dissertation Structure

The proposed dissertation structure consist of six chapters: Introduction, Review of Bimodal Speech Emotion Recognition, Emotion Recognition from Acoustic Features, Combining Acoustic and Linguistic Information at Feature Level, Combining Acoustic and Linguistic Information at Decision Level, and Conclusions.

The first chapter, Introduction, establish presentation of the dissertation. This chapter consist of Background, Motivation, Research Issues, Research Philosophy, Contribution, Organization of Dissertation.

The second chapter is Literature Review. It explains biological and psychological theories of emotion. This chapter consists of Introduction, Emotion Models, Datasets, Features, Classifiers, Fusion Methods, and Summary.

The third chapter, Speech Emotion Recognition Using Acoustic Features, presents evaluations of emotion recognition using acoustic features only including some proposed methods and its limitation. This chapter consists of Introduction, SER using low-level features, SER Using high-level features, combining low- and high-level features, contribution of silent pause frames, and summary.

The fourth chapter, Combining Acoustic and Linguistic Information at Feature Level, features works done on early fusion bimodal SER. This chapter consists of Introduction, Early fusion by Concatenating Feature, Early Fusion by Concatenating Networks, Comparison of Manual Transcription and ASR Outputs, and Summary.

The fifth chapter, Combining Acoustic and Linguistic Information at Decision Level, features works done on late fusion bimodal SER. This chapter consist of Introduction, Data and Feature Sets, Two-stage Bimodal Emotion Recognition, Results and Discussion, and Summary.

The sixth chapter, Conclusions, conclude the dissertation. This chapter consist of summary and future works.

＜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

＜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	