氏名	ATMAJA, Bagus Tris
学位の種類	博士(情報科学)
学位記番号	博情第 445 号
学位授与年月日	令和 3 年 3 月 24 日
論文題目	Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information
論文審査委員	主査 赤木 正人 北陸先端科学技術大学院大学 教授
	党 建武 同 教授
	鵜木 祐史 同 教授
	白井 清昭 同 准教授
	WANG, Longbiao 天津大学 教授
	ARIFIANTO, Dhany Sepuluh Nopember Institute of Technology, Chair

論文の内容の要旨

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.

This study links the current problems in dimensional SER with its potential solutions. The fusion of acoustic and linguistic information fills the gap in dimensional SER. The FL approach improved the performance of unimodal SER significantly. The DL approach improves the FL approach's performance by fusing decisions obtained from the bimodal FL approaches. The results of this research are expected to contribute to gaining better insights for the future strategy in implementing SER, whether to use acoustic-only features (less complex, less accurate), early fusion method (more complex, more accurate), or late-fusion method (most complex, most accurate).

Keywords: dimensional emotion, affective computing, speech emotion recognition, information fusion, acoustic information

論文審査の結果の要旨

本論文は，感情次元（Valence-Arousal-Dominance: VAD）にもとづいて音声中に含まれる感情を自動的に認識するために，音響特徴に加えて言語情報を融合する必要性を調査し，その実現方法を各種提案したうえで，認識率が最大となる最適な手法を提示した研究報告である。

ヒトは，音響的および言語的情報を通じて音声に含まれる感情を伝える。これに従い，感情コンピューティングとして知られるコンピューターによる自動感情認識においても，ヒトの音声感情知覚と同様に，音響および言語にバイモーダル情報を融合することにより，音声感情認識（SER）を実行できる必要がある。本論文では，この目標を達成するために 3 つのサブゴールを設定した。(1) 音響特徴のみを使用する場合の性能向上，(2) 特徴レベルでの音響情報と言語情報の融合，および (3) 決定レベルでの音響情報と言語情報の融合。各サブゴールに対して，以下の結果が得られている。(1) 音響情報のみを使用する場合に感情次元にもとづいた音声認識の能力を最大化することを目的とした課題では，次元 SER のために，音響特徴を表すためのチャンク内での平均と標準偏差の有効性およびのサイレントポーズ領域の重要性を示した。しかし，音響ベースの次元 SER には，次元の一つである Valence の予測精度が低い欠点も見つかった。(2) および (3) では，Valence の予測精度の改善，バイモーダル情報融合の必要性の調査，音響情報と言語情報の融合のための融合フレームワークの評価を目的として，音響言語情報融合のための 2 つ方法（Early Fusion と Late Fusion）を評価した。Early Fusion アプローチでは，音響・言語情報の特徴レベルでの融合，および，音響特徴用ネットワークと言語情報用ネットワー

クの連結を提案した。これらの手法は，音響特徴のみの次元 SER よりも大幅な予測精度の向上を実現した。Late Fusion アプローチでは，音響情報と言語情報が個別にトレーニングされ，結果がサポートベクターマシンによって融合され最終的な認識を得る。評価の結果，Late Fusion アプローチは Early Fusion アプローチよりも優れたパフォーマンスを獲得した。最終結果として，言語情報と音響情報を融合すると Valence 予測のパフォーマンスが向上し，結果として次元 SER の認識精度が向上することを示した。

以上のように，本研究は，次元音声感情認識（SER）のために音響と言語情報を融合する必要性を示し，その実現方法を提案したうえで有効な手法を提示したものであり，学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。