ISS-SP-047

Spectro-Temporal Modulationによる音声感情認識の調査

村上　正悟†　　森田　翔太†
†福山大学工学部情報工学科

1. はじめに

自然なコミュニケーションが可能な機械の実現には，ヒト同士の音声対話と同様に，非言語・パラ言語情報を機械が高精度に認識する必要がある．非言語情報のうち音声感情の認識性能の向上も求められている．音声感情認識では，機械学習やラベル付けに着目した研究が多く，音響特徴量に着目した研究は少ない．音響特徴量としてMFCCがよく知られているが，音声感情認識精度向上には，聴覚特性に基づく音響特徴量の利用が良いと考えられる．聴覚特性との関連も示唆され，音の印象に着目した音響特徴量にスペクトル・時間変調情報（Spectro-Temporal Modulation: STM）[1] がある．

本稿では，STM と深層学習（BiLSTM）による音声感情認識を調査する．

2. VA 空間による感情表出

音声感情認識の感情表出には，怒りや喜びなどの一般的なカテゴリ感情表出と，快－不快のような連続的な変化を前提とする多次元空間の感情表出がある．多次元空間の感情表出の一つに，Valence Arousal Dominanceの3次元のVAD空間がある．ヒトの多様かつ複雑な感情をシステムに組み込むためには，カテゴリ感情表出よりも，連続的に変化する多次元の感情表出が望ましい．VAD空間のDominanceが表出しにくいという報告に基づき，本研究では，Valence ArousalのVA空間を利用する．

3. 評価実験

VA 空間での音声感情認識の精度を比較評価するために，複数の音響特徴量と深層学習による評価実験を行った．深層学習には，Bidirectional LSTM（BiLSTM）を用いた．音響特徴量には，STM 以外に MFCC と聴覚スペクトログラム（HSS）を用いた．感情音声のデータベースには，VAD ラベルが付与された IEMOCAP Database [2]を用いた．感情音声は，学習では 558 音源，評価では学習時に利用していない 16 音源を利用した．サンプリング周波数は，16kHz である．

評価尺度は，Arousal，Valence，Arousal & Valence（VA space）の正解値と予測値の誤差を利用した．ラベル付けされた VA の値はそれぞれ 1～5 である．そのため誤差は 0～4 となる．また，VA space からカテゴリ感情に変換した時の認識性能（Categorical）についても評価した．

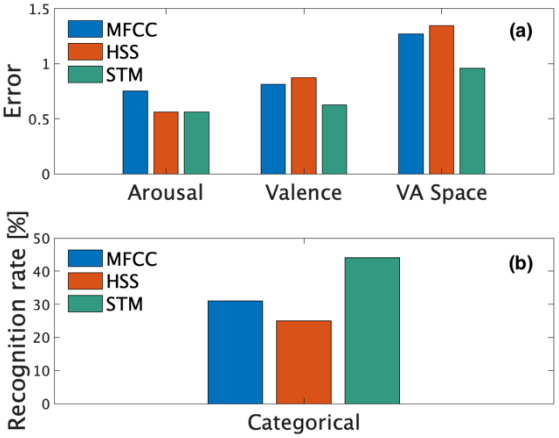

図 1　評価結果　(a) VA 空間での誤差, (b) カテゴリ認識率

4. 評価結果

評価結果を図1に示す．図1（a）はVA空間での平均誤差であり，ArousalではHSSとSTMにおいて誤差0.56でMFCCより良い結果となった．Valenceでは誤差0.63，VA Spaceでは誤差0.96でSTMの誤差が最も小さい結果となった．感情に変換した際の認識精度を示す図1（b）のCategoricalでは，STMの認識率が44%で最も高かった．STMは，全評価項目で最も性能が高い結果を示した．

MFCCやHSSといった一般的な音響特徴量において低い誤差や高い認識率が得られなかった理由は，学習時の音源数が少なかったことが影響している可能性がある．そのような状況下でSTMの性能が最も高かった理由としては，変調情報が感情表出に効果的であった可能性が考えられる．

5. むすび

本研究の結果より，STM が音声感情認識において有効な音響特徴量の一つであるとともに変調情報が重要であることが示唆された．

謝辞

本研究の一部は，科研費・国際共同研究強化（B）（20KK0233）による支援を受けたものである．

参考文献

[1] N. C. Singh, *et al.*, J. Acoust. Soc. Am., Vol. 6, No. 10, pp. 3394-3411, 2003.
[2] C. Busso, *et al.*, Lang. Resour. Eval., Vol. 42, No. 4, pp. 335-359, 2008.