Vol.2019 No.A3-1
2019/3/11

日本語感情音声コーパス JTES を対象とした感情認識の基礎検討

羽田優花[†1]　加藤正治[†1]　小坂哲夫[†1]

日本語感情認識の研究では，フレームごとに計算した特徴量から音全体の特徴を捉えるために統計量の計算をしたものを特徴ベクトルとして認識モデルの学習，認識を行うことが多い．だが，音声認識やその他の分野で，人為的な操作を行っていないデータで学習を行うこと（end-to-end 学習）で認識率の向上が確認されている．そこで本稿ではフレームごとの特徴量から統計量を計算した特徴ベクトルとフレームごとのままの時系列特徴量との比較，検討について報告する．

Basic research on emotional recognition for Japanese emotional speech corpus JTES

Yuka Haneda,[†1] Masaharu Katoh[†1]
and Tetsuo Kosaka[†1]

In Japanese emotion recognition research, recognition models are often recognized by the statistics obtained from frame by frame features in order to capture the characteristics of the whole utterance. However, in the speech recognition and other fields, improvement of the recognition rate has been achieved by end-to-end training. In this paper, we compare the recognition performance between models based on the statistics of whole utterance and those based on the frame-wise features.

†1 山形大学
Yamagata Univercity

1. はじめに

音声認識の技術の発展に伴い，音声対話システムが普及し，一般的に用いられるようになっている．例としては，google の音声検索や iOS の siri などがある．人間の発するメッセージは音声か非音声，言語か非言語の 2 つの軸でおおまかに分類する事が出来る。具体例を挙げると、音声メッセージの場合は言語情報に発話内容、非言語情報には声のトーンや発話速度などがある。非音声メッセージの場合は言語情報であるものに文章や手話、非言語情報であるものに身だしなみや視線や表情、ボディランゲージなどが挙げられる。音声は人間同士の最も基本的なコミュニケーションの手段であるが，そのやり取りには発話内容だけでなく，表情，ボディランゲージ，声のトーン，感情など様々な情報を利用していると言われている．このように人間の音声の持つ情報は発話内容のみではないため，人間と機械の対話の際，情報の伝達に齟齬が生じることがある．そこで，音声から発話内容以外の情報を用いることでより円滑なコミュニケーションが可能になる．音声に含まれる情報として感情があり，音声から発話者の感情を推定することについて取り扱う．現在は多くの対話システムが発話内容のみを用いている．感情認識技術を音声対話システムに応用することにより，より人間同士の対話に近い対話システムを構築することが可能になる．

演技音声の感情認識において識別器としてディープニューラルネットワーク (DNN) を用いることで，サポートベクタマシン (SVM) より認識精度が向上することが分かっている[1)]．日本語でも SVM での感情認識が一般的であった[2)] が最近ではニューラルネットワークを用いた感情認識の研究も行われている[3)4)]．本研究では感情音声データベース Japanese Twitter-based Emotional Speech(JTES)[5)] を対象に，識別器として DNN を用いて感情認識実験を行い，DNN の構造や使用する特徴量などについて検討を行う．JTES を対象とした NN での感情認識の検討は他の研究でもされているが[3)8)]，他システムに組み込むためや予備実験として行われているため，特徴ベクトルの種類などの検討はされていない．

また，従来は時系列の状態の特徴量から統計量 (最大，最小，傾きなど) を計算したものをネットワークへの入力として学習，認識が行われていたが，近年では入力に特徴量のみを用いた認識も検討されている[6)7)]．ニューラルネット (NN) の研究においても人為的な処理を加えたデータを入力とするより，中間の処理も NN に学習させる end-to-end 学習で良い結果が出ている．このことより時系列特徴量から統計量を計算したものと時系列特徴量そのものについて，それぞれを入力とした場合の認識率の比較を行う．

2. 実験概要

感情認識とは，音声に含まれている感情を識別するという技術のことである．人間の感情は発話内容や表情，態度など様々な現れ方をするが，音声の韻律情報にも表れる。これらを特徴量として抽出し，パターン認識を用いて感情の認識を行っていく。本研究では DNN を識別器として用いる．

2.1 特徴量

音声に含まれる感情の特徴は韻律情報 (基本周波数, パワーなど) に現れることが分かっている．発声方法も変化があると言われているため声色の変化を示すメル周波数ケプストラム係数 (MFCC) も使用する．現状，感情認識に明確に必要な特徴量は判明しておらず，関連がありそうな特徴量を全て用いるのが主流となっている．

まず音声が入力されると音声認識と同様に窓関数によってフレームごとに分割され、フレームごとに基本周波数やパワーといった特徴量が計算される。こうして得られた時系列の特徴量は，5 節に示す実験ではここから統計量を計算したものを入力とする．音声に含まれる感情は瞬間ごとの値で表現されるのではなく，発話全体の特徴によって示されると考えられているためである．だが，この統計量が本当に発話全体の特徴を表現するのに十分か不明であるため 6 節に示す提案手法では時系列特徴量のまま入力特徴ベクトルとして用いる．

本研究では INTERSPEECH(2009) の標準セット[9] と The large openSMILE emotion feature set[10] の 2 つの特徴量セットを用いた．それぞれのセットが含む特徴量を表 1 に示す．

2.2 コーパス

本研究では，感情音声コーパス JTES[5] を用いる．JTES は Twitter のつぶやきの中から感情表現語を含む口語的な文章を，音韻や韻律のバランスを考慮し選出したものを用いている．話者は 100 名 (男女各 50 名)，感情は「怒り」「喜び」「悲しみ」「平静」の 4 感情で各感情 50 文，計 20000 発話が用意されている．「自分が意図する感情を機械に伝えるように」発話するようにも指示がされており，人対機械の対話を意識しているというのがこのコーパスの特徴である．人と機械との対話を意識したコーパスであるため、このコーパスで有効な感情認識手法は音声対話システムへの応用が期待できる．また、読み上げるテキストは用意されているが感情強度などの指定はされていないため演技音声に比べ込められた感情にわざとらしさが少なく，自発音声ほどではないが実際私たちが普段発する感情音声に近いデータになっている．

表 1　各セットに含まれる特徴量
Table 1　Features included in each set

構成ファイル	特徴量
INTERSPEECH(2009) 標準セット 16 次元+Δ=32 次元	(RMSenergy，MFCC(1〜12 次元)，零交差率， Voice probability，F0) +Δ
large openSMILE emotion feature set 56 次元+Δ+ΔΔ=168 次元	(LOGenergy，MFCC(0〜12 次元)， melspec(0〜25 次元)，零交差率， Voice probability， F0，F0env， 特定周波数帯のスペクトルエネルギー (0〜250Hz，0〜650Hz，250〜650Hz， 1000〜4000Hz), spectralRollOff (25%，50%，75%，90%)， spectralFlux, spectralCentroid, spectralMaxPos, spectralMinPos) +Δ+ΔΔ

2.3 DNN

DNN は多くの隠れ層を持つようなニューラルネットワークのことである．フレーム毎に状態ラベルを与えて確率的勾配降下法による誤差逆伝搬法によって行われる．このとき，損失関数にはクロスエントロピーを用い，認識時には以下の式で表されるベイズ則に基づきスケーリングを行い，出力確率を求め，音響尤度の計算を行う．

$$p(\boldsymbol{x}js_i) = \frac{p(s_ij\boldsymbol{x})p(\boldsymbol{x})}{p(s_i)}$$

$p(\boldsymbol{x})$ は入力特徴量の生起確率 $p(s_i|\boldsymbol{x})$ は DNN から得られるもの出力で，これを状態生起確率 $p(x)$ で割ったものを出力確率として用いる．

2.4 音声区間検出

5 節で説明する時系列特徴量をネットワークの入力に用いる実験で音声区間検出（VAD: Voice Activity Detection）を用いる．VAD は, 入力音声に対して音声/非音声の判別を行う技術である．音声認識を行う際の前処理でよく用いられ，音声が存在する区間を検出することで高雑音環境下においても高精度な音声認識を可能とする．

3. 時系列特徴量から計算した統計量を学習の入力とした場合の検討

本節ではフレームごとに計算された特徴量から，音声全体の特徴ととるため統計量を計算した特徴ベクトルを用いる．The large openSMILE emotion feature set，6552 次元の特

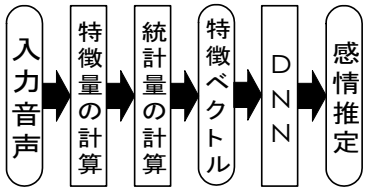

図 1 感情認識実験の流れ
Fig. 1 Flow of emotion recognition experiment

徴量セット特徴量 168 次元 × 統計量 39 次元=6552 次元[10] を用いた．

実験の流れを図 1 に示す．まず入力された音声から声の高さや大きさといった特徴量，特徴量がフレームごとに計算され，時系列で得られる．さらに発話全体の特徴を捉えるために特徴量から各統計量を計算して特徴ベクトルへと変換する．特徴量の抽出から統計量の計算は openEAR[10] を用いて行った．得られた特徴ベクトルを用いて識別器，DNN を用いて感情を識別する．DNN の出力は各感情ごとの尤度であり，尤度が一番高い感情を認識結果とし，正解ラベルと比較して認識率を計算する．

基本の実験条件を表 2 に示す．各実験で特に言及のない場合はこの条件の下で実験を行った．

3.1 dropout 係数とバッチサイズの検討

DNN のパラーメータのうち，dropout 係数とバッチサイズについて検討を行った．実験結果は図 2,3 に示した．

図 2,3 より、dropout なしに比べありの方が全体的に認識率が高い。dropout なしでは学習データの場合バッチサイズ 1024(Accuracy99.24%)，評価データの場合はバッチサイズ 64(Accuracy70.25%) で Accuracy が最大になっている。dropout ありでは学習データの場合バッチサイズ 2048(Accuracy81.38%)，評価データの場合はバッチサイズ 1024(Accuracy73.00%) で Accuracy が最大になっている．この結果より，以降の実験ではバッチサイズ 1024 を採用している．

また学習データ，開発データと評価データの差が小さくなっており、dropout の効果が出ていることがわかる．dropout ありの方がネットワークの汎用性が上がっていると考えら

表 2 基本の実験条件
Table 2 Basic experimental conditions

基本構造	
入力層	6552 次元=(特徴量 56 次元+Δ+ΔΔ)× 統計量 39 次元
中間層	1024×3 層
出力層	4(Neutral,Anger,Joy,Sad)
バッチサイズ	1024
エポック数	validation/main/loss の低下率が 1%未満 もしくはエポック 30 に達すると停止
活性化関数	ReLU
学習法	Adam
dropout 率	入力層 20%, 中間層 50%
使用データ	
学習データ	14400 発話 (40 文 ×4 感情 ×90 話者 (男性 45 話者+女性 45 話者)) このうちランダムで選ばれた 1 割を開発データ, 残り 9 割を学習データとして用いる
評価データ	400 発話 (10 文 ×4 感情 ×10 話者 (男性 5 話者+女性 5 話者))

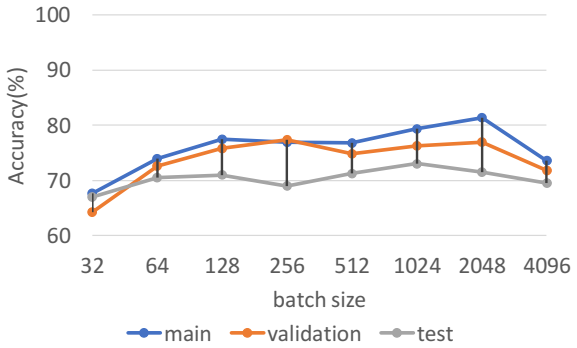

図 2 dropout とバッチサイズの検討 dropout あり
Fig. 2 Consider dropout and batch size with dropout

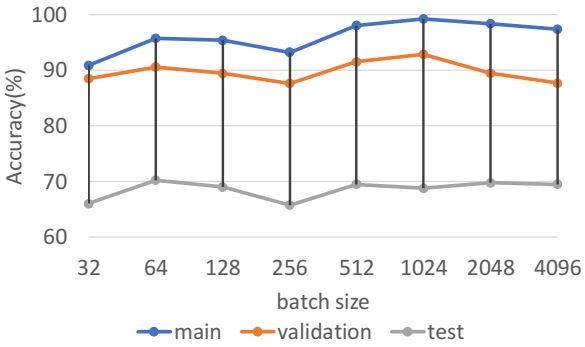

図 3 dropout とバッチサイズの検討 dropout なし
Fig. 3 Consider dropout and batch size No dropout

れ、そのためデータごとの認識率の差が縮まっていると考えられる．

3.2 エポック数についての検討

early stopping という手法でエポック数の決定を行う．学習の進み具合で学習の終了を決定するというものである。ここで，学習の終了条件を変更させて，適切な条件について検討する．以下の 3 つの実験を行った．

(1) エポック数を固定．設定したエポック数は 1〜25.
(2) 開発データの loss 関数の低下度合を基準に early stopping．低下度合は 10，5，2.5，1(従来)，0.75，0.5，0.25，0.1(単位:%)
(3) 開発データの accuracy の上昇度合を基準に early stopping．上昇度合は 10，5，2.5，1，0.75，0.5，0.25，0.1(単位:%)

まず，エポック数を固定した場合の結果を図 4 に示す．

図 4 より，開発データの Accuracy はエポック数を増やすほど良くはなっているが，評価データの認識率は Accuracy，F 値ともにエポック 9 から数値が横這いになっている．このことからエポック数をこれ以上増やしても評価データの認識率の向上はあまり見込めないことがわかる．最も評価データの Accuracy が高かったのがエポック数 11 の際の 73.50(%) である．

次に early stopping 条件についての検討を行う．開発データの loss 関数の低下度合を基

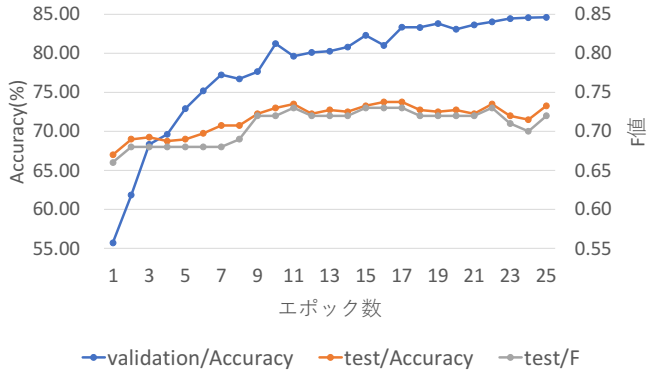

図 4 エポック数を固定した際のエポック数と認識率
Fig. 4 Epoch number and recognition rate when epoch number is fixed

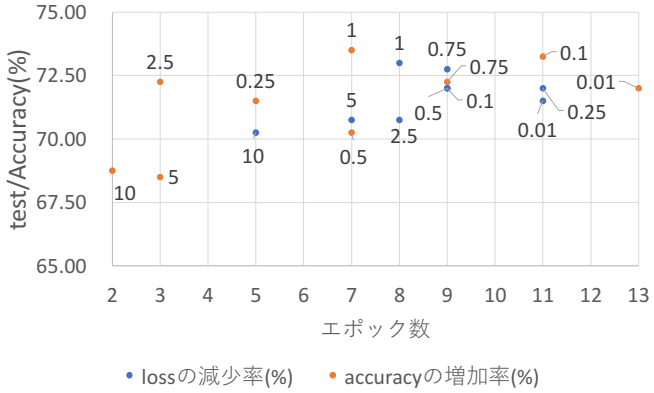

図 5 early stopping を用いた際のエポック数と認識率の散布図
Fig. 5 Scatter diagram of epoch number and recognition rate when using early stopping

準に early stopping を行った場合と accuracy の上昇度合を基準に early stopping を行った場合の結果を epoch 数と評価データの Accuracy の散布図で表したものが図 5 である．

最も高い認識率であるのが，評価データの Accuracy(%) の増加が 1(%) 未満の場合停止

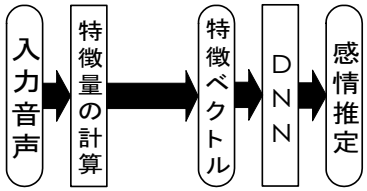

図 6 時系列特徴量を用いた感情認識実験の流れ
Fig. 6 Flow of emotion recognition experiment using time series feature quantity

のもので 73.5%である．従来の条件 (validation/main/loss の低下が 1%未満の場合停止) でも比較的大きな認識率は得られているが validation/main/accuracy の増加を基準とする場合でも最大の認識率を得た 1%や 0.1%などよい認識率を取っているものもある．

4. 時系列特徴量を学習の入力とした場合の検討

前節の実験とは異なり，時系列で得られた特徴量をそのまま特徴ベクトルとして実験を行った。実験の流れは図 6 のとおりである．この実験の問題点として JTES は一発話ごとに一つ感情のラベルが付いているのみで発話区間などの情報はないため，そのままであると無音区間にも正解ラベルが付いた状態で学習，評価を行うことになる．そこで今回の実験では簡易的に評価データにのみであるが VAD を利用し，音声区間のみでの評価を試みた．感情認識，VAD のそれぞれの実験条件は表 3，4 に示す．特徴量は表 1 に示した 32 次元と 168 次元の 2 種類を用い，文脈幅などを変えて実験を行った．実験結果を表 5 に示す．

特徴量でも文脈幅でも値が増えるほど，Accuracy の値は高くなる傾向が確認できる．VAD は用いた際に認識率が上がる場合と下がる，もしくは変わらない場合があった．これについては感情認識モデルの精度によるものだと思われる．図 7 に表 5 に示した場合の評価データの出力尤度の一例を示した．発話内容は「わがまま言ってばっかりだな」である．図中のオレンジの丸で示した個所では尤度の差が発話区間ではっきり出ているが，水色の丸で示した個所は発話区間でもそれほど大きく差が出ていない．この違いから VAD が有効であるモデルとそうでないモデルが出てきたものと考えられる．つまり性能の良い感情認識モデルでは尤度差の大きさにより VAD の働きも兼ねているが，性能の低い感情認識モデルではそれが出来ないため VAD が有効である．また，VAD を行う際，DNN の出力尤度比によって区間検出を行うが尤度に与える重みを変えることで各クラスの敏感度を調節できるため，

表 3 時系列特徴量での感情認識実験条件
Table 3 Experimental condition of emotion recognition with time series feature quantity

基本構造	
中間層	3 層
出力層	4(Neutral，Anger，Joy, Sad)
pre-training	
エポック数	5(1 層目は 10)
バッチサイズ	100
学習法	Contrastiv-Divergence
fine-turning	
エポック数	交差検定によりフレーム認識率の上昇が 1%未満の場合停止
バッチサイズ	256
学習法	Stochastic Gradient Descent
使用データ	
学習データ	JTES14400 発話 (40 文 ×4 感情 ×90 話者 (男性 45 話者＋女性 45 話者))
評価データ	JTES400 発話 (10 文 ×4 感情 ×10 話者 (男性 5 話者＋女性 5 話者))

表 4 VAD モデルの学習条件
Table 4 Learning condition of VAD model

入力層	39 次元 ×41 フレーム
中間層	1024 ユニット ×3 層
出力層	3(無音，雑音，音声)
pre-training	
学習法	Contrastive-Divergence
学習係数	0.4(1 層目は 0.01)
エポック数	5(1 層目は 10)
ミニバッチサイズ	100
モメンタム	0.5〜0.59
L2 正規化係数	0.0002
fine-turning	
学習法	Stochastic Gradient Descent
初期学習係数	0.008
エポック数	交差検定によりフレーム認識率向上が 0.1%未満の場合停止
ミニバッチサイズ	256

表 5　時系列特徴量実験結果
Table 5　Experimental result of time series feature quantity

ネットワーク構造など	VAD 有無	Accuracy(%)
入力 32 次元 ×11 フレーム	なし	64.50
中間層 1024×3 層	あり	67.00
入力 32 次元 ×41 フレーム	なし	66.75
中間層 2048×3 層	あり	66.00
入力 168 次元 ×11 フレーム	なし	67.75
中間層 2048×3 層	あり	69.00
入力 168 次元 ×41 フレーム	なし	68.25
中間層 4096×3 層	あり	68.25

より良い重みを与えることで発声区間検出精度を上げ，その発声区間で評価することで更なる認識率の向上を期待できる．

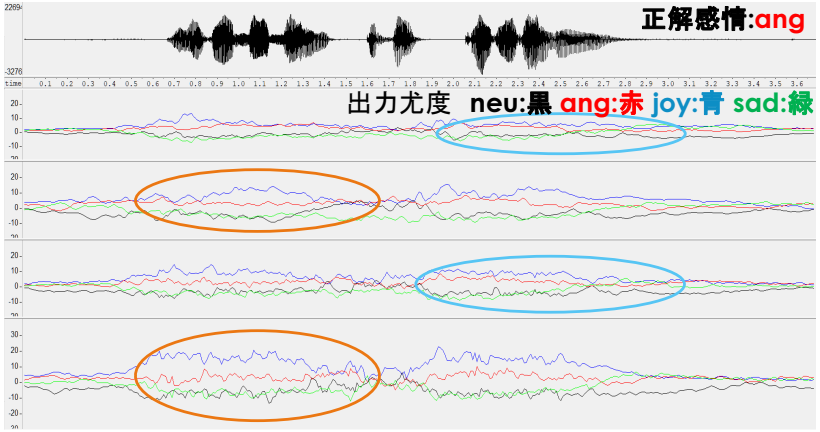

図 7　感情認識の出力尤度
Fig. 7　Output likelihood of emotion recognition

上から順に図 5 と同じ順番で入力 32 次元 ×11 フレーム中間層 1024×3 層，入力 32 次元 ×41 フレーム中間層 2048×3 層，入力 168 次元 ×11 フレーム中間層 2048×3 層，入力 32 次元 ×21 フレーム中間層 4096×3 層の場合である．

前節で最良であった実験結果と本節で最良であった実験結果の比較を図 8 に示す．

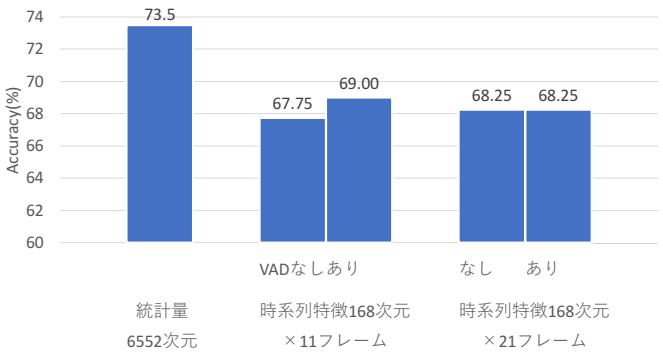

図 8　統計量と時系列特徴量の認識率の比較
Fig. 8　Comparison of recognition rate of statistic quantity and time series feature quantity

前節での時系列特徴量から統計量を計算した特徴ベクトルを用いた場合に比べ，本節での時系列特徴量そのものを特徴ベクトルにした場合に認識率は向上しなかったため，本節の手法と比較すると前節の手法の方が有効であることが分かった．また，本説の手法において認識精度の低い感情認識モデルの場合は VAD を用いることで認識率の向上が確認できたので時系列特徴量に VAD は有効であると言える．

5. 考　　察

時系列特徴量から統計量を計算したものと時系列特徴量で認識率が悪化した原因については大きく二つが考えられる．まず，学習データでの VAD を利用していないことが考えられる．今回は簡易的に評価データにのみ VAD を用いているため，学習データでは発話区間以外にも感情のラベルが付与されている状態である．つまり，発話がない区間では感情を識別できないが，その区間にも正解感情が付いてしまっている状態のデータで感情認識モデルの学習を行った．このことから適切な学習が阻害された可能性がある．統計量では時系列の特徴量をもとにした統計量の計算で発話全体の特徴を丸め込んでいたため大きく影響は出なかったものと考えられる．

次に，VAD モデルとの不適合である．VAD モデルの学習に用いられたのは HAVIC コーパスである．このコーパスはウェブビデオの音声データが中心になっており，雑音や BGM が乗った音声が多く含まれている．一方，感情認識に用いられた JTES コーパスのデータ

コーパス構築のために収録された音声であり，防音室で収録されている．このようなデータの性質の違いから VAD が適切にはたらいていないことが認識率の低さにつながっていると考えられる．

6. 結　　論

本研究では JTES を対象とした感情認識において，DNN への入力ベクトルに時系列特徴量から統計量を計算したものを与える手法と時系列特徴量そのものを入力ベクトルとして与える手法の比較を行った．結果として統計量による方法に比べ，時系列特徴量での認識率の向上は認められなかった．だが，時系列特徴量において精度の低い認識器の場合は VAD で評価データの発話区間を抜き出し認識を行うことは有効であることがわかった．この結果から学習データにも VAD を利用し発話区間のみに感情ラベルを付与したデータでの感情認識モデルの学習や，感情認識に使用した JTES コーパスの特性により適した VAD モデルについて検討することで，認識率の向上を期待できる．また，時系列特徴量を用いる際は感情認識の識別器 DNN を RNN に変更した場合の認識率の検討も行いたい．

参 考 文 献

1) Kun Han, et al. ”Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, in Proc.INTERSPEECH, 223-227.(2014)
2) 武石笑歌 他”エントロピーに基づく音韻・韻律バランス感情依存文の設計と評価”, 電子情報通信学会技術報告書,SP2015-65,pp.33-38(2015-10)
3) 山中麻衣 他,”音響情報と言語情報を用いた協調的発話感情付与に基づく音声対話システムの検討”, 日本音響学会講演論文集, 1-R-30, pp.1037-1038, (2018.9)
4) 真壁大介”自発対話音声を用いた感情認識の研究”, 山形大学修士論文,(2018)
5) 武石笑歌 他”感情音声データベース構築に向けた音韻・韻律バランス感情音声の収録と分析”日本音響通信学会講演論文集,1-R-47(2016-3).
6) Jaebok Kim, et al. ”Towards Speech Emotion Recognition “in the wild” using Aggregated Corpora and Deep Multi-Task Learning”, in Proc.INTERSPEECH, 1113-1117.(2017.8)
7) Ruo Zhang, et al. ”Interaction and Transition Model for Speech Emotion Recognition in Dialogue”, in Proc.INTERSPEECH, 1094-1097.(2017.8)
8) 廣岡信治 他,”要介護者を対象とした音声および感情データベースの構築”, 日本音響学会講演論文集, 2-Q-9, pp.1059-1060, (2018.9)
9) B. Schuller, S. Steidl and A. Batliner, ”The INTERSPEECH 2009 Emotion Challenge,” Proc.IN-TERSPEECH 2009, pp.312-315,(2009).
10) Florian Eyben ”openSMILE-book” , https://www. audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf