感情音声データベース JTES を用いた音声感情認識における特徴量の検討*

☆羽田優花，加藤正治，小坂哲夫 (山形大)

1 はじめに

人間と機械の対話の際，発話内容のみを用いるシステムにおいて情報の伝達に齟齬が生じることがある．そこで，音声から得られる発話内容以外の情報，例えば感情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．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，自然な対話を行うシステムの構築が期待できる．

近年，人間対機械の対話の研究を意図した感情音声コーパスとして Japanese Twitter-based Emotional Speech (JTES) が構築された [1][2]．このコーパスでは感情表現が現れるつぶやきを Twitter から選択し，話者には「自分が意図する感情を機械に伝えるように」発話することが指示されている．

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示されている [3]．日本語でも SVM での感情認識が一般的であったが [1]，近年ニューラルネットワーク (NN) を用いた研究が進展している [4][5][6]．本研究では JTES を対象とした DNN による感情認識において，音響特徴量の検討を行う．

NN を用いた各種研究において，人為的な処理を加えたデータを入力とするより，そのような処理も NN で行う end-to-end モデルで良い結果が得られることが多数報告されている．感情認識において，従来はフレームごとに計算された時系列の Low-Level-Descriptor(LLD)(F0，エネルギー，MFCC など) から固定長のベクトルに落とし込むため，発話全体の統計量 (最大，最小，傾き，偏差など) を求めて特徴ベクトルとするのが一般的であったが，近年では入力に時系列の LLD のみを用いた認識も検討されている [7][8]．JTES を対象とした NN や時系列 LLD を用いた感情認識の検討はいくつか行われているが [4][6]，特徴量の種類の詳細な比較検討はされていない．よって本研究では LLD からの統計量を特徴ベクトルとした場合と，時系列の LLD そのものを特徴ベクトルとした場合について比較検討を行う．

2 実験概要

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，音声の韻律情報やスペクトル情報にも表れる．音声に含まれる音響特徴を抽出し，フィードフォワード型 DNN を識別器として用いて感情認識を行う．認識対象とする感情カテゴリーは怒り，喜び，平静，悲しみの 4 カテゴリーである．本研究においては, kaldi ツールキット [9] を使用して DNN を構築する．本実験では時系列 LLD を特徴量として使用した場合と LLD から発話全体の統計量を求めたものを特徴量として使用した場合の比較を行う．時系列 LLD を特徴量として用いる場合は，無音区間を学習に含めると問題であると考えられるため音声区間検出法を利用した実験も行う．

2.1 感情音声コーパス

本研究では，感情音声コーパス JTES[2] 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．なお，実験を行う際には話者と発話内容について条件が open になるように学習データと評価データの振り分けを行っている．具体的な条件については表 1 に示す．

---

*Investigation on acoustic features in speech emotion recognition using emotion speech database JTES, by Haneda Yuka, Kato Masaharu, Kosaka Tetsuo

Table 1　感情認識実験条件

基本構造	
中間層	256,1024,2046,4096 ユニット ×3〜6 層
出力層	4(Neutral，Anger，Joy, Sad)
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
使用データ	
学習データ	JTES14400 発話 (40 文 ×4 感情 ×(男性 45 話者＋女性 45 話者))
評価データ	JTES400 発話 (10 文 ×4 感情 ×(男性 5 話者＋女性 5 話者))
特徴抽出	
窓幅	25msec
シフト長	10msec
デルタ長	2

2.2　音響特徴量

音声に含まれる感情の特徴は主に韻律情報 (基本周波数, パワーなど) に現れる．音声スペクトルも影響していると考えられるため，メル周波数ケプストラム係数 (MFCC) も使用する．現状，感情認識に必要な特徴量は必ずしも明確でないため，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に分割され，フレームごとに基本周波数やパワー，さらには MFCC といった Low-Level-Descriptor(LLD) が計算される．本研究では INTERSPEECH2009 の標準セット (以下 IS09)[10] と The large openSMILE emotion feature set(以下 large)[11] の 2 つの特徴量セットを用いた．前者は 1 フレーム当たり 32 次元，後者は 168 次元となる．それぞれのセットが含む特徴量を表 2 に示す．

この時系列 LLD に対し発話全体の特徴を表すため各種統計量を計算する．統計量の例としては最大，最小，平均，偏差などがある．IS09 に対する統計量の種類は 12 種であり特徴量の次元は $32 \times 12 = 384$ となる．また large に対する統計量は 39 種で次元数は $168 \times 39 = 6552$ となる．

3.1 節に示す実験では従来法として，以上の方法で得られた発話全体の統計量を特徴量として使用する．一方 3.2 節に示す提案手法での実験では LLD のまま入力特徴として用いる．

2.3　識別器

識別器としては DNN を使用する．構造としては表 1 に示とおり複数の中間層の層数，ユニット数を用い最良のものを結果として示す．学習は pre-training と fine-tuning の 2 段階で行う．pre-training では制限付きボルツマンマシンを使用する．fine-tuning では発話ごとあるいはフレームごとに感情ラベルを与え確率的勾配降下法に基づく誤差逆伝播法で教師付き学習を行う．フレームごとラベルを与える場合は 1 発話について全フレーム同一の感情ラベルを与える．

2.4　音声区間検出

3.2 節で説明する時系列 LLD をネットワークの入力に用いる実験では，無音部を学習に使用すると問題であると考えられるため，音声区間検出技術（VAD: Voice Activity Detection）を利用した実験を行う．本研究では文献 [12] で得られた DNN の VAD モデルを用いて音声/非音声の判別を行った．3 クラス (音声/無音/雑音) 分類を行い，DNN の出力尤度からクラスを決定する．その後音声/非音声区間の継続時間を考慮して，スムージングを行う．範囲の異なる 2 段階の移動平均を用いて過度の湧き出し部分を除去し，この結果を最終的な VAD の結果として使用する．非音声区間と判定された区間は学習時感情クラスのデータから除外する．

3　実験結果

3.1　統計量を用いた感情認識実験

本節では LLD から得られた統計量を用いて認識実験を行った．音声から LLD の抽出, 統計量の計算は openSMILE[13] を用いて行った．特徴量は IS09 特徴セット 384 次元と large 特徴セット 6552 次元の 2 種類である．得られた特徴ベクトルから 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する．評価指標には認識精度 (Accracy) を用いる．

ネットワーク構造は中間層のユニット数 256，1024，2048，4096，層数は 3〜6 層で実験を行い，評価データの Accuracy が高かったものを表 3 に

Table 2　各セットに含まれる LLD

構成ファイル	特徴量
INTERSPEECH(2009) 標準セット (IS09) 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 (large) 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

Table 3　統計量特徴における認識精度

特徴量	中間層	Accuracy(%)
IS09	256×4 層	67.50
large	2048×3 層	69.25

示す．単純に，large 特徴セットの方が次元数が多いため，認識率が IS09 特徴セットを上回るのは想定内と言える．だが次元数が IS09 の約 17 倍であるにもかかわらず Accuracy の上昇は 1.75% であることから特徴としては冗長であると考えられる．

3.2　LLD を用いた感情認識実験

本節では LLD を直接用いた感情認識実験の結果について述べる．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，そのまま用い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．そこで学習データに VAD を用いて非音声区間を検出し non(認識不可) のラベルを付与し，学習を行った．予測感情は全フレームの各クラスの尤度をそれぞれ加算し，尤度が一番高い感情とする．特徴量は表 2 に示した IS09 特徴セットに含まれる LLD32 次元と large 特徴セットに含まれる LLD168 次元の 2 種類を用いて実験を行った．感情の特徴は短時間ではなくある程度の時間幅の音響特徴から抽出できると考えられるため，中間層のユニット数及び層数のほかに文脈幅についても検討を行った．

文脈幅についての例として，VAD ありの large 特徴セットにおいて文脈幅と Accuracy がどのよ

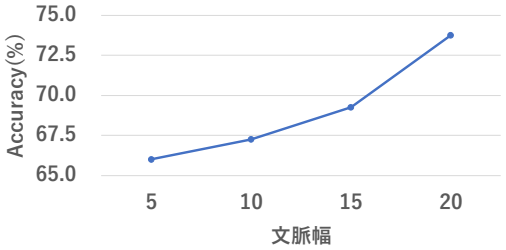

Fig. 1　文脈幅と Accuracy

Table 4　LLD における認識精度

特徴量	文脈幅	中間層	VAD 有無	Accuracy (%)
IS09	20	2048×3 層	なし	66.75
		1024×3 層	あり	70.00
large	20	2048×3 層	なし	67.50
		4096×3 層	あり	73.75

うな関係を示すかを図 1 に示す．この図においては文脈幅とは中心のフレームから前後何フレームを取るかを意味する．これより，文脈幅の値が大きくなるにつれ，Accuracy も上がっていく関係がわかる．感情認識は時間変化で表現されるため，より広い範囲を参照した方が認識率が上がると言える．ただ，文脈幅を増やすと最適な中間層のユニット数も増加する．このため学習にかかる時間も非常に増加する．よって今回は前後 20 フレームまでの検討にとどまった．

各特徴について最良の Accuracy が得られた場合について，文脈幅の値とともに実験結果を表 4 に示す．特徴セットについて IS09 と large では large の方が全体的に精度が高く表れている．前節の結果と変わらず，特徴次元が多い large が高い精度を示しているが，次元数が 5 倍になっているのに対して 1〜3%程度の上昇である．また，

VAD の有無についてそれぞれの特徴セットで 3〜6%程度 VAD ありの方が認識率が上がっていることから学習データに対する VAD は認識に対して有効に働いていることがわかる．

3.3　認識結果のまとめ

3.1 節と 3.2 節の結果をまとめて図 2 に示す．統計量と発話区間の情報を考慮しない時系列 LLD においては統計量の方が高い精度を示しているが，発話区間の情報を考慮した LLD では統計量の精度を上回っており，large 特徴セットにおいては最高の 73.75%の精度を示した．

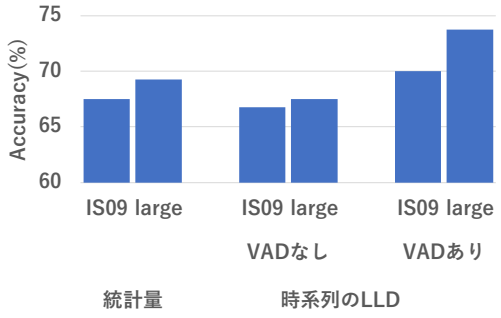

Fig. 2　実験結果まとめ

4　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 LLD から求めた統計量を利用した認識手法と，時系列 LLD そのものを用いる手法の比較を行った．結果として特徴ベクトルに統計量を用いる方法に比べ，発話区間の情報を考慮した時系列 LLD での認識率の向上が認められた．この結果から JTES を用いた感情認識においても，人為的な操作の少ない特徴でより良い認識率を得られることが分かった．

今後の課題として，感情認識においては特徴の時間変化が重要と考えられるため，LSTM など時系列を考慮した識別器に変更した場合の認識率の検討も行なっていく．また，音響的情報の以外の情報，語彙的特徴 (発話内容) や話者の動作，表情など複数のモダリティを用いた感情認識システムの構築も視野に入れて検討を行っていく．

謝辞　感情音声データベース JTES をご提供頂いた東北大能勢准教授に感謝する．本研究の一部は科研費 (課題番号 19K12014) によった．

参考文献

[1] 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)

[2] 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).

[3] Kun Han, et al. ”Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, in Proc.INTERSPEECH, 223-227.(2014)

[4] 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)

[5] 真壁大介, 他”自発対話音声を用いた感情認識の学習データによる検討”情報処理学会東北支部研究会，2017-6-B2-3 (2018.3)

[6] 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)

[7] 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)

[8] Ruo Zhang, et al. ”Interaction and Transition Model for Speech Emotion Recognition in Dialogue”, in Proc.INTERSPEECH, 1094-1097.(2017.8)

[9] ”Kaldi: Deep Neural Networks in Kaldi”http://kaldi-asr.org/doc/dnn.html

[10] B. Schuller, et al.”The INTERSPEECH 2009 Emotion Challenge,” Proc.INTERSPEECH 2009, pp.312-315,(2009).

[11] Florian Eyben”openSMILE-book” , https://www.audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf

[12] 菅郁巳，小坂哲夫，井上雅史, ”DNN を用いた映画の音声区間検出におけるクラス分類の検討”, 日本音響学会秋季講演論文集，1-R-2 (2017.9)

[13] ”openSMILE:” https://www.audeering.com/opensmile/