Vol.2017-GN-100 No.43
Vol.2017-CDS-18 No.43
Vol.2017-DCC-15 No.43
2017/1/21

話者の感情表示コンテンツに向けた感情音声認識技術に関する研究

坂野太亮[†1]　木川貴博[†1]　杉本雅則[†2]
楠　房子[†3]　稲垣成哲[†4]　溝口　博[†1]

概要：人との関わりにおいて、他者の感情を推定することは重要な役割を果たす。感情を推定するための有力な手がかりとして口調や笑声、泣き声などの音声がある。機械が音声から感情を推定できれば、人との関わりを支援することが可能となる。そこで、著者らは音声から感情を推定し、その感情を視覚化するコンテンツの作成を目指している。本講演では、感情の現れた音声の 1 つである笑声の認識技術について報告する。

キーワード：深層学習，笑声認識，メル周波数ケプストラム分析

Study on Emotion Recognition using Emotional speech toward Digital Contents Expressing Speaker's Emotion

TAISUKE SAKANO[†1]　TAKAHIRO KIGAWA[†1]　MASANORI SUGIMOTO[†2]
FUSAKO KUSUNOKI[†3]　SHIGENORI INAGAKI[†4]　HIROSHI MIZOGUCHI[†1]

Abstract: With a goal of developing digital contents expressing speaker's emotion, we try to recognize emotion using emotional speech. As the first step to recognize emotion, we start to recognize laughing voice for recognizing emotion of pleasure. A previous study start to recognize a laughing voice based on the periodic waveform of the laughing voice. However, a problem arises when recognizing laughing voices through their waveform, since it is possible to have false positive results. To overcome this problem, we proposed laughing voice recognition method that incorporates a voice-likeness feature. To improve the recognition success rate, we try to incorporate deep learning into the laughing voice recognition method we proposed. To confirm the efficacy of the laughing voice recognition method incorporating deep learning, we compared with proposed method, discrimination success results improved by 3%. Therefore, the proposal method can be assumed to recognize an effective means of laughing voice.

Keywords: Deep learning, Laughing voice recognition, Mel-frequency cepstrum analysis

1. はじめに

人が社会的生活をおくるうえで，誰しも他者と関わりをもっている．他者と友好的な関わりをもつためには，コミュニケーションをとることが必要不可欠である[1]．コミュニケーションの手段の中でも会話や通話など，話をすることは有効な手段である[2].

しかしながら聴覚障碍者など，耳の聞こえに障碍をもつ人は口調などに現れた他者の感情や状態を理解することが難しい．そのため，話をすることによるコミュニケーションの利点を十分に活用できないという問題が考えられる．この問題を解決するため，著者らは音声から推定した話者の感情を視覚化するコンテンツの作成を目指している．このコンテンツを開発するためには，音声からの感情推定が必要である．

人の感情は、「喜怒哀楽」に分けることができる[3]．人の感情を推定するためには笑声，泣き声，怒号など，喜怒哀楽が現れた人の様子を認識することが必要である．そこで著者らはその第一歩として，「喜」の感情を推定するために有力な手がかりとなる笑声の認識に関する研究に着手した．

笑声の認識にはさまざまなものが提案されてきた[4]．喉頭部マイクロホンから会話中の音声を取得し笑声を認識する爆笑計[5]や，笑声の振幅やその変化の特徴から笑声を認識するシステム[6]などが報告されている．これらの手法では，入力された音の波形が周期的であるという笑声の特徴を用いて笑声認識を行っている．そのため，周期的な波形であれば，笑声以外の音までも笑声として誤認識をしてしまうという課題があった[4][5][6].

この課題を解決するために著者らは「声らしさの特徴」を認識に用いる手法[7]，声らしさの特徴に加え周期的波形の特徴をも認識に併用する手法[8]を提案した．そして，これら手法のさらなる識別性能向上を目指し，現在，深層学習を取り入れた笑声認識に取り組んでいる．本論文では，

†1 東京理科大学
Tokyo University of Science
†2 北海道大学
Hokkaido University
†3 多摩美術大学
Tama Art University
†4 神戸大学
Kobe University

Vol.2017-GN-100 No.43
Vol.2017-CDS-18 No.43
Vol.2017-DCC-15 No.43
2017/1/21

実データを用いて行った深層学習を取り入れた笑声認識の有効性確認実験について述べる．

2. 深層学習を用いた笑声認識

声らしさの特徴を認識に用いる手法[7]，声らしさの特徴に加え周期的波形の特徴をも認識に併用する手法[8]では，笑声，笑声以外の音声についてメル周波数ケプストラム変換[9]を行っている．メル周波数ケプストラム変換により，時間-メル周波数に対する対数振幅強さが得られる．そして，時間-メル周波数平面上に表される対数振幅強さを画像のように見立て，画像の特徴量抽出手法である高次局所自己相関を用いて特徴量抽出を行っている．

これらの手法の有効性が確認されていることから，時間-メル周波数に対する対数振幅強さのグラフについて深層学習を用いて学習させる笑声認識が可能ではないかと著者らは考えた．具体的な深層学習手法としては，画像認識で成果をあげている深層学習である畳み込みニューラルネットワーク[10]を用いる．畳み込みニューラルネットワークには，鳥や虫など多数の一般物体画像をあらかじめ学習させたものを使用し，その畳み込みニューラルネットに笑声，笑声以外の音声を再学習させる．再学習により笑声，笑声以外の音声を認識するための最適化を行う．この章ではメル周波数ケプストラム変換について，畳み込みニューラルネットワークを用いた深層学習について述べる．

2.1 メル周波数ケプストラム変換

メル周波数ケプストラム変換は音声認識などに用いられている音声処理手法である．メル周波数ケプストラム変換によれば音声信号の低周波数領域の形状を顕著に表すことができる．

処理の概要を図 1 に示す．音声について短時間フーリエ変換を行い，音声の持つ時間に対する振幅の情報を周波数と時間に対する振幅強さの情報に変換する．変換された周波数と時間に対する振幅強さの情報について，振幅強さを対数化，周波数を人の聴覚特性を表すスケールであるメルスケールに変換する．その後，離散フーリエ変換に周波数成分を低周波数領域に集中させる工夫をした離散コサイン変換を行う．

2.2 畳み込みニューラルネットワーク

畳み込みニューラルネットワークは画像認識の分野でよく用いられる深層学習手法であり，人の視覚野にある受容野をモデル化している．

畳み込みニューラルネットワークは，データを入力する入力層，入力に対して畳み込み処理を行い，入力の特徴を表す 2 次元の特徴マップを得る畳み込み層，畳み込み層から出力された特徴マップのデータサイズを縮小するプーリング層，2 次元の情報である特徴マップを 1 次元に展開する全結合層，認識結果を出力する出力層，が多層に連結した構成となっている．これらの層により，特徴抽出，認識を行う．

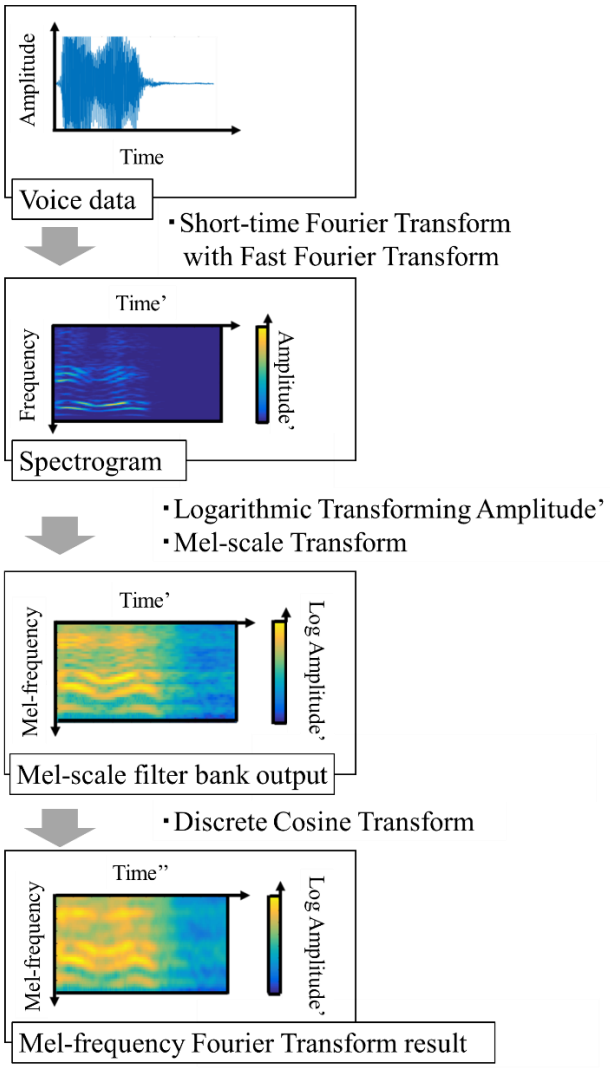

図 1　メル周波数ケプストラム変換の概要
Figure 1　Overview of Mel-frequency cepstrum transform

3. 有効性確認実験

深層学習を用いた笑声認識の有効性を確認するため，笑声，笑声以外の音声の実データを学習させ，それぞれの分類を行った．この章では，使用した畳み込みニューラルネットワークの構成，音声データ，深層学習により学習させたデータ，学習結果について述べる．

3.1 使用した畳み込みニューラルネットワークの構成

使用した畳み込みニューラルネットワークの構成を図 2 に示す．使用した畳み込みネットワークは，入力層，5 層の畳み込み層，3 層のプーリング層，3 層の全結合層，出力層から構成されている．

3.2 使用した音声データ

使用した笑声のデータは，笑声 2 周期分の長さとし，1 つの笑声とした．笑声以外の音声についてはハ行の音を含む単語の発声を 1 つの笑声以外の音声とした．　笑声 192 デ

Vol.2017-GN-100 No.43
Vol.2017-CDS-18 No.43
Vol.2017-DCC-15 No.43
2017/1/21

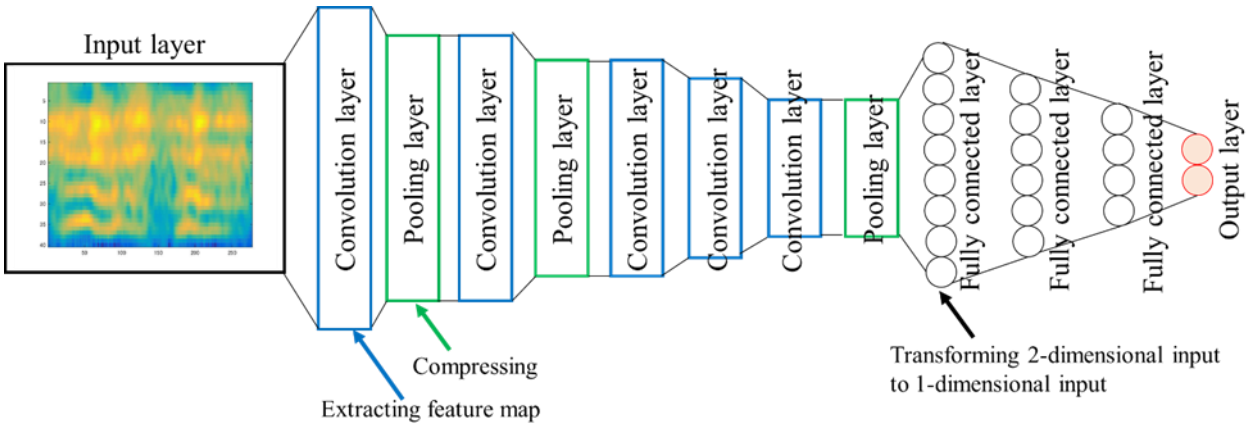

図 2　使用したニューラルネットワークの構成

Figure 2　Constitution of using CNN

ータ、笑声以外の音声 192 データを用意した．笑声のデータについてはラジオ番組から，笑声以外の音声データについては情報・システム研究機構 国立情報学研究所 音声資源コンソーシアムの音声コーパスである重点領域研究「音声言語」・試験研究「音声 DB」連続音声データベースから取得した．笑声データの長さについて，最大は 0.58 秒，最小は 0.15 秒，平均は 0.33 秒である．笑声以外の音声について，最大は 2.51 秒，最小は 1.43 秒，平均は 1.72 秒である．ただし，笑声以外の音声には発声のない区間が含まれている．データのサンプリング周波数は 16000Hz である．離散化の分解能は 16bit である．

3.3 深層学習により学習させたデータ

深層学習により学習させたデータは，笑声，笑声以外の音声についてメル周波数ケプストラム変換を行い，得られた時間（単位はケフレンシー）-メル周波数-対数振幅強さのグラフである．学習させたデータの一例を図 3 に示す．学習させたグラフには時間とメル周波数軸のメモリも含まれている．画像のサイズは縦 656 ピクセル，横 865 ピクセルである．画像の深度は RGB 各 8bit ずつの 24bit/pixel である．

3.4 学習結果

笑声，笑声以外の音声についてメル周波数ケプストラム変換を行い，得られたグラフを学習させた．学習により，笑声のデータ群と笑声以外の音声のデータ群とを分類する．このとき，笑声のデータであり，笑声のデータ群が多数含まれる領域にあるデータの総数と笑声以外の音声であり，笑声以外の音声のデータ群が多数含まれる領域にあるデータの総数を足し合わせ，笑声，笑声以外の音声の全データ数で割ったものを分類正解率と定義する．式に分類正解率の定義を示す．深層学習により学習させた結果，分類正解率は 100%となった．このことから，深層学習を用いた笑声認識の有効性が確認された．

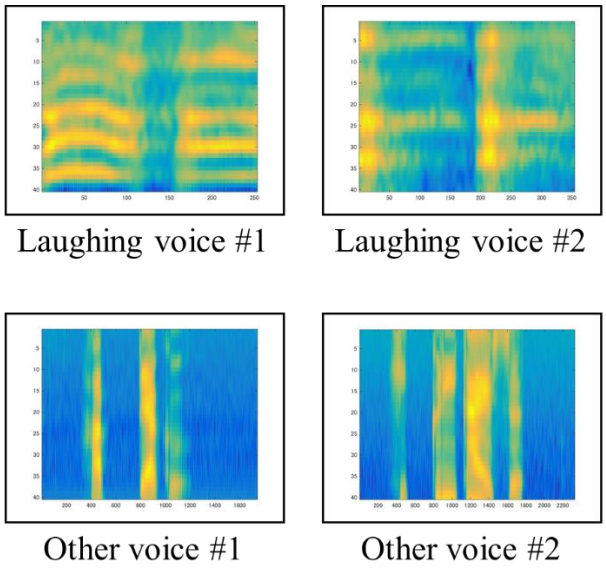

図 3　学習させた笑声，笑声以外の音声のグラフの例

Figure 3　Example of learning data
(laughing voice and other voice)

4. おわりに

声らしさの特徴を認識に用いる手法[7]，声らしさの特徴に加え周期的波形の特徴をも認識に併用する手法[8]のさらなる識別性能向上のために，深層学習を取り入れた笑声認識に取り組んだ．

深層学習を取り入れた笑声認識では，まず，笑声，笑声以外の音声についてメル周波数ケプストラム変換を行い，時間-メル周波数に対する対数振幅強さのグラフを取得する．次に取得したグラフについて畳み込みニューラルネットワーク用いて学習させる．

実データを用いて学習を行ったところ，分類正解率が 100%となった．これにより深層学習を用いた笑声認識の有効性が確認された．

Vol.2017-GN-100 No.43
Vol.2017-CDS-18 No.43
Vol.2017-DCC-15 No.43
2017/1/21

謝辞 本研究の一部は，JSPS 科研費 JP26560129，JP15H02936 の助成を受けた．音声データの一部に，情報・システム研究機構 国立情報学研究所 音声資源コンソーシアムの音声コーパスである重点領域研究「音声言語」・試験研究「音声 DB」連続音声データベースを使用した．記して謝意を表す．

参考文献

[1] 町田佳世子．コミュニケーション能力，ストレス対処，意欲の関連．札幌市立大学研究論文集. 2006, vol.3, no.1, p. 35-44.
[2] 田中真詞，川端豪．音声対話を通した共同作業タスクの検討．電子情報通信学会技術研究報告. Speech. 1995, vol.95, no.123, p.89-94.
[3] 佐藤高弘，中川匡弘．フラクタル次元解析を用いた感情の定量化手法―感性フラクタル次元解析法―．“電子情報通信学会技術研究報告.HIP. 2002, vol.102, no.534, p.13-18.
[4] 久間秀樹，高橋勇作，福岡 久雄，玄行 照朗，皆尾 登志美．ホビーロボットを用いた高齢者介護施設における「笑い」の定量的評価方法．日本笑い学会笑い学研究. 2010, no.17, p.50-60.
[5] 松村雅史，辻龍之介．笑い声の無拘束・長時間モニタリング－爆笑計－．電子情報通信学会技術研究報告.SP，音声. 2005, vol.105, no.105, p.7-12.
[6] 田中爽太，鈴木健嗣．笑い声に応答するロボットのための笑い声認識システムの開発．日本機械学会ロボティクス・メカトロニクス講演会 2015(ROBOMEC2015)講演論文集. 2015, p.1A1-S06.
[7] 坂野太亮，木川貴博，竹村裕，溝口博．機械による人の感情・状態推定に向けた非言語認識に関する研究―基礎的検討としての笑声認識―．日本機械学会ロボティクス・メカトロニクス講演会 2016(ROBOMEC2016)講演論文集. 2016, p.2A1-09b7.
[8] Sakano,T. Kigawa,T. Sugimoto,M. Kusunoki,F. Inagaki,S. and Mizoguchi,H.. Laughing Voice Recognition Using Periodic Waveforms and Voice-likeness Features --Towerd Advanced Human-machine--. Proceedings of the IEEE International Conference on Robotics and Biomimetics. 2016, p.964-969.
[9] 荒木雅弘．イラストで学ぶ音声認識．講談社, 2015.
[10] 山下隆義．イラストで学ぶディープラーニング．講談社, 2016.