

テキストと音声のマルチモーダルな感情推定

森川寛也^{†1} 三橋晟^{†2} 河合継^{†3} 能勢隆^{†4} 千葉祐弥^{†5}

概要: 本論文では、音響情報による感情識別の結果と言語情報による感情識別の結果を結合した感情認識の手法を提案する。具体的には音声を音響的な情報と言語的な情報に分けてそれぞれの感情識別器を作り、2つの識別結果を結合して感情識別を行うという方法である。音響情報の感情識別器については、文章を単語ごとに区切り1単語1セルとしてLSTMへ入力し学習させる方法と、区切らずに学習させる方法とを比較した。結果としては現在の時点では区切らないほうが精度が高いが、今後データ量の増加に伴い区切った方がよい精度を出す予想できる結果となった。また言語と音響の識別結果を結合したモデルではテストデータに対して、78.2%の精度で感情を識別することができた。これは音響情報のみと言語情報のみの識別結果よりも精度が高かった。今後の課題としてさらに精度を上げていくには大規模コーパスを作成か、データ拡張によってデータを増やしていく必要があることが分かった。

キーワード: 感情認識, 深層学習, 音声, 自然言語処理

Multimodal Emotion Recognition Using Lingual and Acoustic Feature

HIROYA MORIKAWA^{†1} SHO MITUHASHI^{†2} KEI KAWAI^{†3}
TAKASHI NOSE^{†4} YUYA TIBA^{†5}

Abstract: In this paper, we propose the method of emotion recognition that combines the emotion identified result of auditory and language information. Specifically, speech is divided into auditory and linguistic information to create an emotion identifier machine for each. In terms of emotion identifier of the auditory information, we separated the sentence by each word and compared the two method. One learned by in-putting one word per cell to LSTM. The other learned without separation. As a result, for now, the accuracy is higher when the sentences are not separated. However, as the amount of data is expected to increase in the near future, we assume that it would be better to separate the sentences for an improved accuracy. Moreover, in a model that combined the identifying result of language and auditory, we were able to identify the emotion with accuracy of 78.2% for test data. The number was higher than the identifying results of auditory only or language only. This leads to the need for creating a large corpus or increasing the amount of data by data expansion in order to further improve accuracy as a future task.

Keywords: Emotion recognition, Deep learning, Voice, Natural language processing

1. はじめに

近年、機械の計算スピードやメモリの目ざましい向上により機械はさまざまな分野でこれまでにない成果を残すようになった。音声認識の分野においては非常に精度がよく、日常生活で見ない日は無いほどその技術が活用されている。例えば、GoogleHome [1] やAmazonEcho [2] などのAIスピーカーにそれらの技術が使われている。これらは音声を入力にして音楽を流したり、質問に答えたりと執事のような役割を果たすものとなっている。このように人間同士による対話のような機械のインターフェースが一般化していき、人と機械は今後密接な関係になると思われる。そのような関係となるために必要な課題の1つとして、機械が人の気持ちを理解するという課題がある。その課題を解決し人の気持ちを理解することができれば、人に寄

り添った対応ができるようになる。音声によるインターフェースは既に実用化されており、また私達是对話において音声の言語情報からだけでなく、非言語情報・パラ言語情報からも発話者の意図、気持ちを自然とくみ取っている。このようなことから、音声を用いた感情識別の研究は数多く行われており非常に重要な分野となっている。

音声による感情識別の研究においては、MFCC やピッチといった音響特徴量をニューラルネットなどの識別器を用いてクラスタリングする手法が多く取られていた[3]。しかし、近年ではコンピュータの計算速度の向上から、音響特徴量を用いずEnd-to-Endで行おうとする研究も行われている[4]。また、マルチモーダルな感情推定の研究も行われており、例えばビデオから映像・言語・音響の情報を入力としての感情推定が行われている[5]。マルチモーダル

^{†1} ^{†2} ^{†3} (株)クリスタルメソッド

Crystal-method Co., Ltd.

^{†4} ^{†5} 東北大学大学院工学研究科・工学部

School of Engineering, Tohoku University

にすることで識別精度の向上は期待できるが、その向上率は劇的なものでないことが多い[6]. その要因の一つとしてマルチモーダルにする時に間違った単位で結合してしまっているのではないかとすることが挙げられる. 例えば、音響と言語の結合においてフレーム単位でお互いの特徴を結合したとしても、音響にとっては効果があるが言語にとってはそれによって意味のある特徴を手に入れることは難しい. マルチモーダルな感情推定においては特徴抽出機と各特徴量結合の2段階の結合が行われている[7]. このような感情推定においては、基本的に音声と顔による組み合わせが多い. こうした中、怒りについて言語と音響の関係性や組み合わせの研究がある[8]. また[9] では言語と音響の組み合わせについて上記の特徴抽出と各特徴量の結合の2つの工程を一回の学習で行っている. これによって最適な重みを見つけることができる. この手法が有効になっていることの要因の一つとしては大量の学習データセットがあるということが挙げられる. しかし、言語によってはそのようなデータセットが必ずしもあるというわけではない. 現状の日本語感情識別におけるデータセットでは自由度の高すぎる[9] のようなネットワークは厳しいと思われる. 日本語のマルチモーダル感情認識については言語情報と音響情報が一致した場合には一致したものを結合結果とし、もし相違があれば音声の識別結果を優先して感情識別をするというようなルールベースな手法で両者の結合を試みていた[10].

そこで本論文では、まず音響情報による感情識別のところでは意味の最小単位である単語単位での特徴を識別に加味することを試みた. 次に言語情報による感情識別であるが、こちらは4つの基本感情にラベル付してあるデータセットがなかったので、弊社の研究チームでデータを作成し、さらに精度向上のためデータ拡張を行った. 最後に音響情報と言語情報の結合のやり方として、音響情報の感情識別器と言語情報の感情識別器をそれぞれ別に作った上で、それぞれの感情識別結果の各感情の確率を入力にしたネットワークを作った. ネットワークの概要は1に示す. ここでは、音響の感情推定にLSTMを用いたものを表した. こうすることによって、マルチモーダル化における結合単位が統一され、またより少ないデータで音響と言語の最適な結合方法を見つけ、音響もしくは言語のみの識別よりも精度が良くなることが予想される. 実際、音響と言語を組み合わせた感情識別モデルがテストデータに対して75.5%という結果を出した. これは音響単体、言語単体で学習した結果よりも精度が高く、結合によって性能が向上していることが確認された.

本論文の流れとしては、第2章でまず音響情報を単語に分割する手法を説明し、分割したものと、分割していない識別モデルを提案する. 次に、言語については既存の4つの基本感情における既存のデータセットがなかったので

データ・セット作成の手順から、データの拡張法について解説する. そして章の最後に音響情報と言語情報を組み合わせた感情識別のデータ、モデルについて提案する. 第3章でそれらの結果をまとめた上で考察し、第4章で研究を進めていくうえで出てきた課題を記述し、最後に第5章で本論文のまとめ、そして今後の目標についてまとめる.

2. 提案手法

音響による感情認識と言語による感情認識をそれぞれ2つにわけて学習を行い、最後にそれぞれの学習器の出力結果を結合する. 本章では、音響と言語のデータ、音響による単語ごとに分ける時と分けない時の学習手法、次に言語による学習手法、それから音響と言語の結合モデルについて述べる.

2.1 音響による感情認識

2.1.1 データ

音響データにはJTES(Japanese Twitter-based Emotional Speech)[11]を使用している. JTESとは、100人の男女(各50名ずつ)が、angry/joy/neutral/sadの4つの感情で、各感情50種類ずつの文章(発話)を読み上げた音声データセットである. Twitterの呟きをベースとしており、感情ラベルの付与、音韻・韻律のバランスが取れた文選択などの特徴がある. このデータを、モデルの学習に用いる学習データ、モデルのハイパーパラメータを決めるためのバリデーションデータ、モデルの検証のためのテストデータの3つに分けた. 学習データ、バリデーションデータ、テストデータでは発話者と発話内容はかぶっていない(テキスト、話者オープン). 拡張データは、東北大学の能勢准教授、千葉助教から頂いたもので、テラーメイド音声合成[12]と呼ばれる手法で生成されたものである. テラーメイド音声合成とは、一度音声合成した音声に対して、そのイントネーション・リズム等を後から変更したものである. ここでは、JTESの話者100名(男女各50名)の話者の音響モデルを元にして、4つの各感情でJNAS(Japanese-Newspaper-Article-Sentences)[13]の文章100文を合成したものをを用いる. JNASとは毎日新聞記事文を読み上げた音声コーパスである. これによって感情識別可能な語彙数を増加させた. 音声データを分割する前処理の段階でうまくいかなかったデータがあったためデータの合計数がわずかに少なくなっている. 音響データの内訳については表1に示す.

表 1 音響学習器で使ったデータ

	話者数	発話数	計
学習データ	80名 (男女各40名)	4感情×30発話	9594発話
拡張データ	80名 (男女各40名)	4感情×100発話	31996発話
バリデーションデータ	10名 (男女各5名)	4感情×10発話	400発話
テストデータ	10名 (男女各5名)	4感情×10発話	400発話

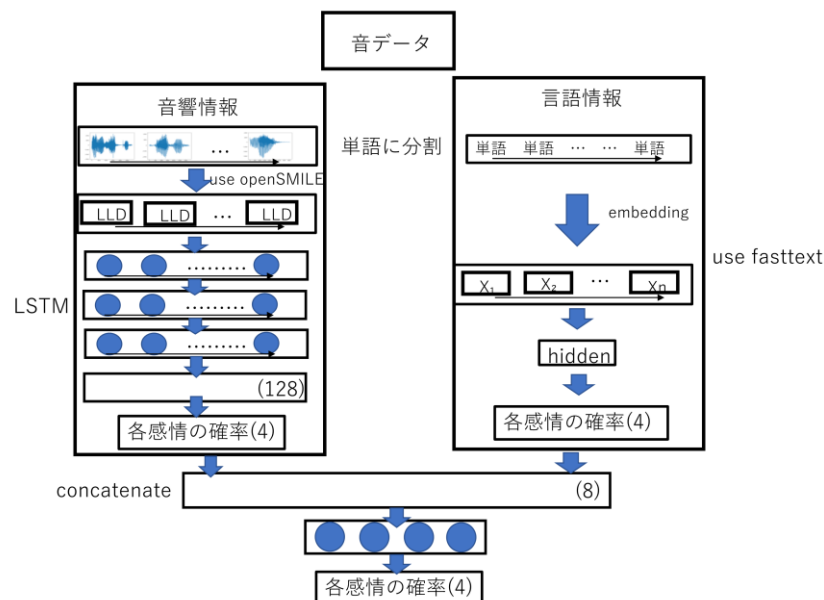


図1 音声の音響情報と言語情報から感情推定を行うネットワーク図

2.1.2 音声の分割

音響データをに単語ごとに分割する。分割することによって文章単位としてだけでなく、単語単位での感情認識が可能になる。方法としては以下のような手順で行った。

(1) 形態素解析システムのMeCab[14] (辞書はipadic-NEologd[15]) を用いて発話文章(テキスト) を単語ごとに分割した。

(2) 連続音声認識ソフトウェアのJulius[16] のセグメンテーションキットを使い発話時間を手に入れ、それに基づき音声データを単語ごとに分割する。Juliusのセグメンテーションキットを用いると、書き下し(Transcription) ファイルと音声ファイルを与えることで、書き下しの内容に従って音素のアライメントを行うことができる。

2.1.3 音響の特徴量抽出

音響情報を、openSMILE[17] を利用して音響データを384次元の特徴量ベクトルにする。openSMILEとは、音声データから感情を認識するために、様々な音響データの特徴量を出力できるツールである。openSMILEではどういいう特徴量を出力するかという設定をコンフィグファイルで行う。任意のコンフィグファイルを選択することが出来るが、本論文ではIS09 emotion.conf[18] を選択した。openSMILEで取得可能な特徴量を表2に示す。

このように、音響情報をそのまま扱うのではなく特徴量として扱う。これらの特徴量をLow Level Descriptors (LLD) と呼び、LLD から平均値や分散、傾き、最大値や最小値などの各種統計量を計算する。音声の中の単語の数を n 個とすると、単語ごとのLLD時系列に対してそれぞれ m 個の統計量を計算し、入力された音声は $n \times m$ 次元

の特徴量ベクトルへと変換される。この特徴量ベクトルに対して、0 から1 の範囲で正規化を行った。

表2 openSMILEで取得可能な特徴量

特徴	説明
RMSenergy	音量の二乗平均平方根値
MFCC	1次～12次メル周波数ケプストラム係数
F0	基本周波数
voiceProb	その時点での音が声である確率
pcm_zcr	波形のゼロ交差率

2.1.4 FCL-baseの音響感情識別

単語に分割しない感情識別のモデルとして、一文をそのまま感情識別したものはFCL(Fully-Connected-Layer)-baseのものにした。一文を384次元のベクトルに直して、3つの全結合層を挟んで出力層から各感情の確率を算出した。全結合層の活性化関数はReLU、出力層はsoftmax関数と設定した。epoch数は100、最適化方法はAdgrad、損失関数はcategorical crossentropy、評価関数はaccuracyを用いた。各種設定は表3に示す。

また、categorical crossentropy(loss) は以下の式で表す。 N がデータの数を表し、 M はクラスの数を表す。ここでは $M=4$ となる。

$$loss = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad (1)$$

表 3 FCL-base ネットワークの各種設定

Layer_Name	Content
入力層	特徴量ベクトル (384)
全結合層_1	128
dropout 率_1	0.2
活性化関数_1	ReLU
全結合層_2	128
dropout 率_2	0.5
活性化関数_2	ReLU
全結合層_3	128
dropout 率_3	0.5
活性化関数_3	ReLU
全結合層	4
活性化関数	Softmax

2.1.5 LSTM-base の音響感情識別

Long-Short-Term Memory) [19] を用いた。LSTMはRNN (Recurrent Neural Network) の拡張として1995 年に誕生した、時系列データに対するモデルである。従来のRNN では学習できなかった長期依存を学習可能であるということが、LSTM の大きな特徴としてある。1 単語 (384 次元の特徴量ベクトル) を1 セルとし、3 つのLSTM-Layerの後に、全結合層の出力層から各感情の確率を算出した、LSTM-Layer の活性化関数はReLU、出力層の活性化関数にsoftmax 関数を用いることで、確率の形にして出力した。epoch 数、最適化関数、損失関数、評価関数は学習器(2.1)に使ったものと等しいものを使用した。これらの深層学習のモデルにはKeras[a]を用い、パラメータは実験的に設定した。各種設定は表4 に示す。

表 4 LSTM-base ネットワークの各種設定

Layer_Name	Content
入力層	単語数 (time_step), 特徴量ベクトル (384)
LSTM 層_1	(time_step,128)
dropout 率_1	0.2
活性化関数_1	ReLU
LSTM 層_2	(time_step,128)
dropout 率_2	0.5
活性化関数_2	ReLU
LSTM 層_3	(time_step,128)
dropout 率_3	0.5
活性化関数_3	ReLU
全結合層	4
活性化関数	Softmax

2.2 言語による感情認識

2.2.1 データ

日本語感情のラベル付きの文章データセットは公開されていなかったため、弊社研究チームで作成することになった。言語データは、千葉大学3 人会話コーパス[20]、名大会話コーパス[21]、Twitter コーパス[22]、livedoor ニュースコーパス[23] から集めた。コーパスとは、テキストや発話を大規模に集めてデータベース化した言語資料である。集めた言語データに対して、以下のようにして感情のラベル付けを行った。

(1) 文章のネガポジ判定をGoogle Cloud Natural Language API[24] のanalyzeSentiment メソッド[b]を使用して行った。これは、指定されたテキストを調べてそのテキストの背景にある感情的な考え方を分析するものである。記述されたテキストの感情極性を判断する。

(2) 極性判定されたものを参考に、テキストを4 つの感情とその他に手作業で分けた。

しかしながら、手作業で仕分けるため十分な量の感情ラベル付きデータセットを集めることが困難であった。そこで集めたラベル付きデータの学習データを元にしてデータ拡張を行った。データ拡張手法は、学習データを英語とドイツ語に一旦翻訳したものを再度日本語に翻訳しなおす手法である。言語による言葉や文法の違いから翻訳するときにある言葉が省かれたり、違う言い回しになったりすることがある。その性質を利用し翻訳したものを日本語に再翻訳すると、その過程で言い回しが変化しデータの拡張が期待できる。今回、翻訳にはGoogle Cloud Translate API[25]を使用した。この手法によって学習データと同じ意味だが、微妙に言い回しが異なるデータを生成した。そしてテストデータには、学習データと同じ手法で集めたデータ (4 つの感情それぞれに100 文で計400 文) を用いた。言語データの内訳については表5 に表す。

表 5 言語学習器で使用したデータ

	ang	joy	neu	sad	計
学習データ	1950	1950	1950	1950	7800
拡張データ	3900	3900	3900	3900	15600
テストデータ	100	100	100	100	400

a) <https://keras.io/ja/>

b) <https://cloud.google.com/natural-language/docs/analyzing-sentiment>

2.2.2 言語の感情識別手法

この3種の拡張したデータを用いてそれぞれの精度を比較した。識別手法は、fasttext[26][27][c]のテキスト分類機能を用いて、1文における各感情の確率を求めた。fasttextはFacebook AI Research[d]製のライブラリで、分かち書きされたファイルを入力すると、単語のベクトル表現の生成とテキストの分類が高速に行えるライブラリである。fasttextによるテキスト分類は出力層の活性化関数にsoftmax関数が使われているので、各感情の確率を出力することが出来る。epoch数は20で、精度はaccuracyで表した。

2.3 音響と言語の感情認識の結合

2.3.1 データ

音響と言語の感情認識の結合の際に用いたデータは、2.1で用いたデータから拡張データを除いたものである。データの内訳については表6に示した。

表6 音響と言語の識別結果から感情を推定するときのデータ

	話者数	発話数	計
学習データ	80名(男女各40名)	4感情×30発話	9594発話
バリデーションデータ	10名(男女各5名)	4感情×10発話	400発話
テストデータ	10名(男女各5名)	4感情×10発話	400発話

2.3.2 音響と言語の結合ネットワーク

音響と言語情報の結合のさせ方としては以下のようにした。音響のみで学習させた学習器(2.1)と、言語のみで学習させた学習器(2.2)それぞれの重み、バイアスを固定した。この状態で各識別器からアウトプットされる各感情に対する確率を入力にして、一層の全結合層を挟みsoftmax関数を掛けることによって各感情における推定確率を出力する。これによって音響と言語情報の結合のみを学習することができる。音響情報と言語情報の結合の式は以下に示す。P(X),P(Y)は言語、音響情報のみから得られる4つの感情に対する確率であり、Wはそれぞれに対する重み、bはバイアスである。

$$A_p = \sum_{j=1}^4 P(X_j) \times W_{X_{jp}} + \sum_{k=1}^4 P(Y_k) \times W_{Y_{kp}} + b \quad (2)$$

$$S_p = \frac{\exp(A_p)}{\sum_{i=1}^4 \exp(A_i)} \quad (3)$$

epoch数、最適化関数、損失関数、評価関数は(2.1)に使ったものと等しいものを使用した。

3. 結果と考察

まず、音響のみの識別結果は図2に示す。FCL-base・LSTM-baseそれぞれの識別結果はepoch数の中で、validationデータに対してlossが最も低い時の重みを使用した結果である。どちらのネットワークもデータ拡張後の方が精度が高く、データ拡張が有効であったことが確認される。また、両者の差はデータ拡張の前後、どちらもFCL-baseの方が良い。しかし、データ拡張後はお互いの差が少なくなっていることが見受けられる。ここからLSTM-baseのネットワークの方が精度の向上が良いことがわかる。この結果より、データが今後増えていくと、LSTM-baseのネットワークの方がFCL-baseのネットワークより良くなることが予想される。これはLSTMがデータの並びも学習するので、通常の学習よりも精度が良くなるまでに大量の学習データが必要であるためだと考えられる。

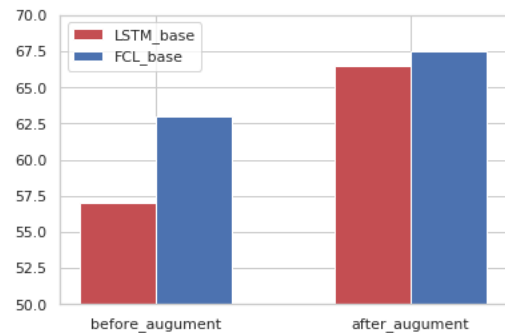


図2 音響のLSTM-baseとFCL-baseの精度比較

次に、言語のみの感情識別結果は、学習データ拡張前後の精度を比較したものを表7に示した。

表7 言語のデータ拡張の精度

	拡張前	拡張後
accuracy(%)	46.3	47.5

拡張前よりも拡張後の精度が良くなっているという結果になった。それぞれの語彙数について表8に記述した。

表8 学習データと拡張データそれぞれの語彙数

	拡張前	拡張後
語彙数	8667	12513

拡張データは学習データにない語彙が再翻訳される間に増えていき、そのため拡張前よりも精度が向上したのではないかと考えられる。例えば元の文章が「未だにあの巻を読み返すのは心苦しいもん」という文章が再翻訳されると

c) <https://fasttext.cc/>

d) <https://research.fb.com/category/facebook-ai-research/>

「そのボリュームをまだ読み返すのは難しいです」というようになる。ここでは「ボリューム」「難しい」という語彙が増えている。精度が向上した拡張データだが、生成された文章が元の文章とどれほど似ているニュアンスで違う文章になっているという検証を以下のようにして行った。オリジナルのデータ (Ori), 英語を経由して再翻訳したデータ (Eng), ドイツ語を経由して再翻訳したデータ (Ger) という3つのデータから、2つを選びDoc2Vecによりベクトル化したもののcos類似度を比較した結果の平均値を表9に示した。

表 9 Doc2Vec による比較の平均値

感情 \ 比較言語	Ori - Eng	Ori - Ger	Eng - Ger
ang	0.45	0.44	0.70
joy	0.47	0.45	0.73
neu	0.43	0.41	0.73
sad	0.47	0.45	0.73

Doc2Vec とは, Le らが発表したParagraph Vector[28] という手法で, 任意の長さの文書をベクトル化し, 文章やテキストの分散表現を得ることができる技術である。

Doc2Vecの学習モデルにはdmpv(Distributed Memory Paragraph Vector) を用いた。dmpv とは周辺の単語ベクトルと文章ID を入力にして出力を予測単語として学習する手法である。これにより, 文章における文脈が加味された分散表現を表すことができる。dmpv の学習データは, 元データと拡張データの文章である。この文章の分散表現同士におけるcos類似度を出すことによって似ている文章, 似ていない文章を数値として表すことができる。文章の類似度は-1から1の間の値で表され, 1 に近づくほど2つの文章は似ており, -1 に近づくほど2つの文章は似ていないものと判定する。文章の類似度におけるデータの例を表10に示した。

表 10 Doc2Vec の例

	文章	類似度
Original English	満員電車で隣のお姉さんが寄りかかってきて重かったです 次の妹は混乱した列車に傾いてきて重い	0.24
Original German	明日から夏日続きで昨日買ったコートが早速お蔵入り 昨日朝に買って, 夏の日に持ち歩いたコートはすぐ店に来ます	0.46
English German	犬はそのような日にのみ眠らないだろう そのような日に犬は寝ない	0.72

最後に, 音響のみによる識別結果, 言語のみによる識別結果と音響と言語の結合による識別結果の比較を行う。音響と言語の結合ネットワークの識別結果は音響による識別と同様にepoch 数の中でvalidation に対するloss の最も低いものを使用する。音響・言語・音響+言語の精度をそれぞれ表11 に表した。

表 11 音響単体, 言語単体, 音響言語結合による精度

	sound		text	sound+text	
(sound-network)	lstm	fcl		lstm	fcl
accuracy	66.5%	67.5%	47.5%	75.5%	78.2%

音響・音響+言語におけるネットワークについては音響ネットワークがFCL-base のものとLSTM-base のものそれぞれに対しての精度を表した。各感情の具体的な識別結果を表12 で表している。こちらでは音響ネットワークはFCL-base のものを使用している。表11 から, 音響のみによる識別よりも, 音響と言語を組み合わせた識別のほうが高い精度が出ていることがわかる。3つの識別器を比較すると, 音響+言語, 音響のみ, 言語のみの順に精度が高くなっていた。言語情報の感情識別は他と比較すると低い精度になっているが, これは学習データが感情を特定するにはデータが少なく精度を上げることが困難であったからであると考えられる。

4. 課題

今回の研究で音声に関してはデータ数の不足により精度がそれほど上がらないという問題や, 用いるデータによって精度が大きく影響される問題があり, 次に言語においてはang/joy/neu/sad の日本語文章データセットが公開されていない問題, fasttext によるテキスト分類では学習時に使われている単語しか対応できなく未知語に対応できない問題などが浮かび上がってきた。また今回は行わなかったが, 言語と音響を組み合わせるやり方としてはもう1つ考えられる。それは音声データから音響と言語の両方の特徴量を抽出し, ひとつの識別器で両方の情報を加味した識別結果を出力するこるやり方である。しかし, 現状では学習に用いることのできる音声データが不足していて言語情報の識別に必要な語彙数を確保できないという問題がある。これらの問題を解決するには規格化された大規模コーパスの作成か, 音声データの拡張によって大量の感情を含む文章を生成する必要がある。

5. おわりに

人間の対話は発言の内容はもちろんのこと声のトーンなどの音響情報も加味して行われており, 音声対話システムにおいても内容と声の両方を加味することで感情を認識して対話を行うことが期待される。そこで本論文では, 音声の要素を音響情報と言語情報それぞれで感情識別を行い, それらの結果を結合して音響と言語の両方を用いた感情識別の精度を検証した。結果として音響と言語を結合した識別結果が最も高い精度を出すというものとなった。今後, 弊社研究チームで様々な感情音声データの入手, あるいは

表 12 音響単体・言語単体・音響言語結合による感情識別

actual\predict	sound(fcl)				text				sound(fcl)+text			
	ang	joy	neu	sad	ang	joy	neu	sad	ang	joy	neu	sad
ang	65	26	7	2	54	6	12	28	91	0	7	2
joy	26	52	10	12	2	78	14	6	9	76	9	6
neu	0	5	74	21	6	16	54	24	3	7	61	29
sad	0	4	17	79	20	8	12	60	7	3	5	85

データ拡張の開発を行い前章の課題を解決していきたい。

参考文献

- [1] Google(日本): Google Home, 入手先
(<https://store.google.com/jp/product/googlehome>)
(参照2018-08-05)
- [2] Amazon(日本): Amazon Echo, 入手先
(<http://amzn.asia/gn17e2K>) (参照2018-08-05)
- [3] Pierre-Yves Oudeyer: "The production and recognition of emotions inspeech: features and algorithms", International Journal of Human-Computer Studies, Volume 62, Issue 3, March 2005, Pages 423
- [4] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, Stefanos Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network" Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5200-5204, 2016.
- [5] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu, "Recognizing Emotions in Video Using Multimodal DNN Feature Fusion" Association for Computational Linguistics, pp. 11-19, 2018
- [6] Sidney D' Mello,acqueline Kory: "Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies", in Proceedings of the 14th ACM international conference on Multimodal interaction. ACM, 2012, pp. 3138.
- [7] Soujanya Poria,Iti Chaturvedi,Erik Cambria,Amir Hussain "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis" 2016 IEEE 16th International Conference on Data Mining (ICDM) 12-15 Dec. 2016
- [8] Tim Polzehl,Alexander Schmittb,Florian Metzger,Michael Wagnerd: "Anger Recognition in Speech Using Acoustic and Linguistic Cues" Volume 53, Issues 910, NovemberDecember 2011, Pages 1198-1209
- [9] Yue Gu, Shuhong Chen, Ivan Marsic: "DEEP MODAL LEARNING FOR EMOTION RECOGNITION IN SPOKEN LANGUAGE", arXiv:1802.08332
- [10] 鈴木基之: "音声に含まれる感情の認識", 日本音響学会誌, Vol.71, No.9, pp.484-489 (2015).
- [11] 相澤佳孝, 加藤正治, 小阪哲夫ほか: "感情音声データベースJTES を用いた感情音声認識におけるモデル適応の性能向上の検討", 研究報告音声言語情報処理, Vol.2017-SLP-119, No.7, pp.1-6 (2018).
- [12] 山田修平, 能勢隆, 伊藤彰則: "テラーメイド音声合成のための差分特徴量を用いたDNNに基づくF0制御", 日本音響学会研究発表会論文集(CD-ROM), Vol.2017 Issue. 春, pp.ROMBUNNO.1-Q-21 (2017).
- [13] 国立情報学研究所: JNAS - 音声資源コンソーシアム, 入手先(http://www.ipsj.or.jp/journal/submit/manual/j_manual.html) (参照2018-08-05) .
- [14] 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 入手先(<http://taku910.github.io/mecab/>) (参照2018-08-05) .
- [15] Toshinori Sato: mecab-ipadic-NEologd, GitHub 入手先(<https://github.com/neologd/mecab-ipadic-neologd>) (参照2018-08-05) .
- [16] Julius development team: 大語彙連続音声認識エンジンJulius, OSDN 入手先(<http://julius.osdn.jp/>) (参照2018-08-05) .
- [17] Eyben, F., Wllmer, M., Schuller, B.; openSMILE The Munich Versatile and Fast Open-Source Audio Feature Extractor, the 9th ACM Multimedia, pp1459-1462. (2010).
- [18] Schuller, B.,Steidl, S.,Batliner, A.: The INTERSPEECH 2009 Emotion Challenge, Proc. Interspeech 2009, pp.312-315 (2009)
- [19] Hochreiter, S. and Schmidhuber, J.: Long short-term

memory, Neural computation, Vol.9, No.8, pp.1735-1780 (1997).

[20] 国立情報学研究所：Chiba3Party -

音声資源コンソーシアム, 入手先

(<http://research.nii.ac.jp/src/Chiba3Party.html>) (参照2018-08-05) .

[21] 国立国語研究所：名大会話コーパス入手先

(<https://mmsrv.ninjal.ac.jp/nucc/>) (参照2018-08-05) .

[22] 大崎彩葉, 唐口翔平, 大迫拓矢："Twitter 日本語形態素解

析のためのコーパス構築", 言語処理学会第22回年次大会発表論文誌, pp.16-19 (2016).

[23] ロンウィット：livedoor ニュースコーパス入手先

(<https://www.rondhuit.com/download.html#ldcc>) (参照2018-08-05) .

[24] Google Cloud：Cloud Natural Language API ,

Google Cloud 入手先(<https://cloud.google.com/natural-language/>) (参照2018-08-05) .

[25] Google Cloud: Cloud Translate API ,Google Cloud 入手

先(<https://cloud.google.com/translate/?hl=ja/>)??.

[26] Bojanowski, P.,Grave, E.,Joulin, A., et al.: Enriching Word Vectors with Subword Information, arXiv preprint (2016)

[27] Joulin, A.,Grave, E., Bojanowski, P. et al.: arXiv preprint (2016)

[28] Le, Q V. and Mikolov, T.: Distributed Representations of Sentences and Document , Proceedings of the 31st International Conference on Machine Learning, Vol.32, No.2, pp.1188-1196 (2014).