

# ボトルネック特徴量を用いた感情音声の認識

向原 康平<sup>1,a)</sup> サクリアニ サクティ<sup>1</sup> グラム ニュービック<sup>1</sup> 戸田 智基<sup>1</sup> 中村 哲<sup>1</sup>

**概要:** 音声認識システムは一般に広く使われるようになっており、感情のこもった音声を自動認識する機会が多くなってきた。しかし、感情の揺らぎは音声に影響を与え、平静音声を想定している通常の音声認識システムでは認識精度の低下を引き起こす。本研究では感情音声の入力に対する認識精度向上のため、ディープニューラルネットワークを用いて作成するボトルネック特徴量に着目する。ボトルネック特徴量とは、隠れ層のノード数を少なくしたボトルネック構造のニューラルネットワークから抽出する特徴量である。ボトルネック特徴量は特徴量強調が行われ、感情音声のゆらぎに左右されない音素の本質的な成分を抽出できると考えられる。ボトルネック特徴量を用いたモデルを感情音声の認識に用いた結果、通常の音声認識モデル、感情に対して適応学習を施したモデルと比べて音声認識精度の向上が確認できた。

**キーワード:** 感情音声認識, ボトルネック特徴量, ディープニューラルネットワーク

## Bottleneck Features for Emotional Speech Recognition

KOHEI MUKAIHARA<sup>1,a)</sup> SAKRIANI SAKTI<sup>1</sup> GRAHAM NEUBIG<sup>1</sup> TOMOKI TODA<sup>1</sup> SATOSHI NAKAMURA<sup>1</sup>

**Abstract:** Automatic speech recognition (ASR) system is used for emotional speech. However emotion influence speech signal. Therefore emotional speech degrades ASR quality. In this study, we focus on bottleneck features for emotional speech recognition. The bottleneck features are made by deep neural network hidden layer which has small number nodes than other layer. We think bottleneck structure can extract features and bottleneck features represent phoneme essential features. By using bottleneck features for emotional speech recognition, we confirm improvement results compared with emotion adaptation model and normal model.

**Keywords:** emotional speech recognition, bottleneck features, deep neural network

### 1. はじめに

音声認識は技術的な発展からめざましい進歩 [1] を遂げているが、その認識精度が発話の環境や性質に強く依存することは依然として大きな問題である。例えば雑音のある野外や残響の発生するホール、人間が感情を込めた音声 [2] を入力する場合などがあげられる。このような問題に対して、周囲の環境による影響を軽減するための雑音除去や残響除去 [3] は多く行われているものの、人間の感情に対応した音声認識システムはまだ少ない。

通常、音声認識では平静状態で発声された音声情報と言語情報の関係をモデリングするため、感情音声をそのまま音声認識システムに入力した場合、認識精度は低下する [4]。従来、感情音声の認識には音声の感情状態や変動に対して適応学習手法が用いられてきた。音声特徴量に対して適応学習する手法 [5] や発話の内容に対して適応学習する手法 [6] である。適応学習手法は適応のターゲットとなる感情とテストデータの感情が一致している場合、認識精度の向上が見込めるが、一致しない場合は認識精度の向上は見込めない。また感情に関して適応モデルを作成する場合、モデルの数は定義する感情の個数必要になる。

本研究では従来法の問題点を解決するために、ディープニューラルネットワークを用いたボトルネック特徴量 [7]

<sup>1</sup> 奈良先端科学技術大学院大学 情報科学研究科  
Graduate School of Information Science, Nara Institute of  
Science and Technology (NAIST), Japan.

<sup>a)</sup> mukaihara.kohei.me4@is.naist.jp

の抽出を感情音声の認識に用いる手法を提案する。ボトルネック特徴量はニューラルネットワークの中間層のユニット数を少なく抑えるボトルネック構造のネットワークから抽出される。ボトルネック構造の中間層で抽出している特徴は入力特徴量を次元圧縮し、入力の代表的な特徴を表現した値になっていると考えられる。本研究では感情音声をネットワークに入力し、音素を出力のターゲットとして学習を行う。そのため感情による音素のゆらぎに影響の少ない特徴量になって出力されることが期待される。従来法と比較すると、感情を含む音声をすべて使ってネットワークを作成することができるため、感情に合わせてモデルを作る必要がなくどの感情にも対応できる。

実験では通常モデル、感情適応モデル、ボトルネック特徴量モデルのそれぞれに対して感情音声を認識させた。その結果、ボトルネック特徴量モデルはモデルを複数作ることなく、他のモデルと比べて感情音声の認識精度の改善が確認できた。

## 2. GMM/HMM 音声認識

本研究では感情音声に対して音響モデルの対応を行うことで認識精度向上を図る。本節ではその時に用いられる、音響モデルとその適応学習について説明を行う。

### 2.1 音声認識の定式化

音声認識は、入力音声系列から得られるフレーム単位の特徴量ベクトル系列  $X$  を用いて単語列  $W$  を求める問題としてとらえることができる。この問題は以下のように定式化することができる [8]。

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X) \quad (1)$$

この問題における  $P(W|X)$  は直接推定する代わりに、ベイジの定理を用いて以下のように変形することが一般的である。

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(W)P(X|W)}{P(X)} \quad (2)$$

この時、 $P(X)$  は音声認識の結果に関わらず音声特徴量により一定となり、分母にある  $P(X)$  は結果には影響を与えないため、取り除くことができる。

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W)P(X|W) \quad (3)$$

与えられた特徴量ベクトル系列に対して式 (3) を満たす  $W$  を決定するのが音声認識の計算になる。音声認識において  $P(W)$  を言語モデル、 $P(X|W)$  を音響モデルと呼ぶ。本研究では音響モデルを取り扱うことで感情音声の認識精度を向上させる。

### 2.2 音響モデル

音響モデルは音声認識において  $P(X|W)$  で表現される

ことを前節で述べた。これは与えられた単語  $W$  から特徴量ベクトル系列  $X$  が生成される確率をモデル化したものである [9]。通常、単語  $W$  から直接  $X$  を推定されるのではなく単語を構成する音素単位でモデル化されるのが一般的である。単語と音素列の関係は発音辞書と呼ばれる辞書によって定義されている。音響モデルは音素と音声特徴量を隠れマルコフモデル (HMM: Hidden Markov Model) を用いてモデル化する。

HMM の各状態に対する出力確率はその音素に対する音声特徴量の分布を決定するモデルで決まる。このモデル化は混合ガウス分布 (GMM: Gaussian Mixture Model) やディープニューラルネットワークなどで行われる [10]。

### 2.3 MLLR 適応

音声認識には発話者のごく少量の音声を用いて発話者の発声の音響的な特徴をシステムに適応させる話者適応化 [11] という技術がある。感情音声においても適応学習が使われる [12]。ここではその代表例の一つである最尤回帰 (MLLR: Maximum Likelihood Linear Regression) 法について説明を行う。MLLR 法は少ない学習データ量でも高い性能が期待できるため話者適応において広く用いられてきた。MLLR 法は HMM の持つ各ガウス分布の平均ベクトル  $\mu = (\mu_1, \dots, \mu_n)$  を変換させることで行われる。式 (4) に  $\mu$  に関するアフィン変換の式を示す。この時、 $n$  は特徴ベクトルの次元数を表している。

$$\hat{\mu} = A\mu + b \quad (4)$$

ここで  $A$  は  $n \times n$  の行列であり、 $b$  は次元数  $n$  のベクトルである。この  $A, b$  を尤度最大化基準で推定する。全てのガウス分布で一つの  $A, b$  を共有する場合は入力データ量が少ない場合でもモデルの変化が見込め、性能改善が期待できる。

## 3. 提案手法

本研究ではボトルネック構造のディープニューラルネットワークから抽出するボトルネック特徴量を用いて、感情音声を入力とした場合の認識精度向上を図る。本章では提案手法を構成する要素についてそれぞれ説明を行う。まずネットワークから得た特徴量を音声認識に用いる Tandem アプローチ [13] について説明を行う。その後ボトルネック特徴量について説明を行い、最後にネットワークがどのように学習されるか説明を行う。

### 3.1 Tandem アプローチ

Tandem アプローチはディープラーニングが音声認識に適用される前から提案されてきた MLP (Multi Layer Perceptron) と GMM の複合アプローチである [14]。音声認識ではより高い認識性能を実現するために特徴量選択や特徴量抽出の工夫がされてきた。Tandem アプローチもそ

の一つであり MLP を特徴量抽出のためのモデルとして用いる。Tandem アプローチでは通常の入力ベクトル  $\mathbf{x}_t$  は MLP の出力ベクトルを正規化したベクトル  $\Psi(\mathbf{y}_t)$  に変換される。この  $\Psi(\mathbf{y}_t)$  は GMM によってモデル化され、元の特徴ベクトル  $\mathbf{x}_t$  の出力分布  $P(\mathbf{x}_t|q_t)$  は  $P(\Psi(\mathbf{y}_t)|q_t)$  に比例するとして利用される。このときの MLP は入力特徴量の音素を学習するように設計されるのが一般的である。Tandem アプローチの場合は補助的に識別問題を解くように学習することでそのネットワークを特徴量抽出の部品とする。このアプローチの優れている点はディープラーニングによる学習を用いてよりよい特徴量を発見することができる点である。本研究では Tandem アプローチをディープニューラルネットワークに適用した。

### 3.2 ボトルネック特徴量

本研究で用いるボトルネック特徴量はディープニューラルネットワークの中間層のうち、1層を他の層のノード数よりも少なくした層から得ることのできる特徴量である。一般的にボトルネック構造のネットワークはクラス分類に必要な情報を適切に低次元の特徴量に変換することができるといわれている。ボトルネック特徴量を用いたニューラルネットワークの代表的なものはオートエンコーダである。オートエンコーダはボトルネック構造であるネットワークによって自身の情報を低い次元に落とすことと出力層に自身を写像する。この時ボトルネック中間層には入力情報をよく表現できるような学習がなされる。音声認識にボトルネック構成を用いる場合、入力は特徴量ベクトル、出力は音素ラベルの形が多い。この場合、ボトルネック特徴量として期待されるのは音素の本質的な情報を再現可能な形で表現できる特徴量であり、ネットワークはそのように学習されることが期待される [15][16]。本実験で用いるボトルネック構造のディープニューラルネットワークの構造を図 1 に示す。

ボトルネック構造による学習は pre-training と fine-tuning に分けられる。pre-training では入力層から順にボトルネック中間層に向けてオートエンコーダを用いた学習を行う。pre-training が終了するとボトルネック中間層から HMM 状態を表現している出力層までをつなげて学習する、fine-tuning を行う。この時、出力層は HMM 状態数の数だけノードを持っており、入力音声に対応する音素を出力するように学習が行われる。

## 4. 実験

### 4.1 実験設定

異なる構造を持つ音響モデルに対して感情音声を入力しそれぞれの認識精度を確認する。本実験では日本語話し言葉コーパス (Corpus of Spontaneous Japanese, CSJ) のコアデータ (約 45 時間, 50 万語) をもとにして音響モデル

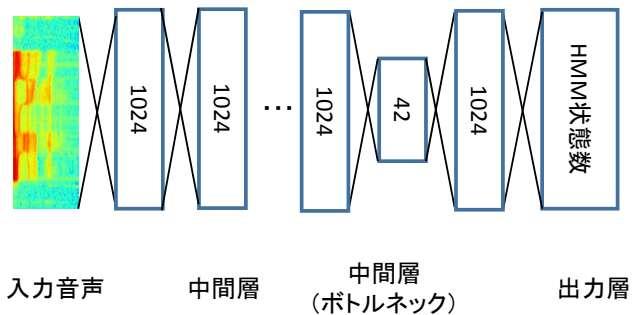


図 1 ボトルネック構造ディープニューラルネットワーク

Fig. 1 The bottleneck structure of deep neural network.

を学習した。具体的には 3 通りのモデルを比較する。

#### CSJ モデル

CSJ のみを学習データとして用いた音響モデル。

#### CSJ/MLLR モデル

実験時、入力される感情音声と同じ種類の感情を適応学習するモデル。

#### CSJ/BNF モデル

感情音声を学習データとして、ボトルネック構造のディープニューラルネットワークをからボトルネック特徴量を抽出する Tandem アプローチを適用したモデル。

それぞれの音声認識システムに対して、感情音声を入力し各感情に対して有効かどうか確認する。

本研究において、感情音声を用いたテスト・学習には感情評定値付きオンラインゲーム音声チャットコーパス (OGVC: Online gaming voice chat corpus with emotional label) を用いた [17][18]。OGVC はオンラインゲーム中のプレイヤー同士の音声チャットの発話をもとに収録されており、自発対話音声と演技音声の 2 種類で構成されている。本研究では、演技音声を感情音声として用いた。演技音声は自発対話音声の発話テキストの中から感情が表出されている発話を選択し、プロの俳優 4 名 (男性 2 人, 女性 2 人) が感情を込めて発声した音声を収録している。発話には受容 (ACC), 怒り (ANG), 期待 (ANT), 嫌悪 (DIS), 恐怖 (FEA), 喜び (JOY), 悲しみ (SAD), 驚き (SUR), 以上 8 種類の感情ラベルが付与されている。発話者は感情の強度を変えて 4 種類の音声を収録した。強度はそれぞれ感情を含まない平静状態 (レベル 0) と弱 (レベル 1), 中 (レベル 2), 強 (レベル 3) である。発話者 1 名につき 664 発話, 計 2656 発話を収録している。

本実験では OGVC に定義されている 8 種類の感情音声と平静状態の音声を実験データとする。感情音声に関しては同じ発話内容で感情の強度のレベルが 4 段階あるので、強度のレベルごとに感情の影響をそれぞれ確認する。テストデータは男性 1 人, 女性 1 人がそれぞれ収録した音声を合わせて 1 つのデータとする。各感情それぞれ 40 発話をテストデータとして用いる。

ボトルネック構造ディープニューラルネットワークの学

表 1 実験条件-CSJ モデル

Table 1 Experimental conditions - CSJ model.

音声特徴量	LDA-MLLT (40 次元)
音響モデル	CSJ コーパス
言語モデル	OGVC 664 発話
発音辞書	OGVC 367 語

表 2 実験条件-CSJ/MLLR モデル

Table 2 Experimental conditions - CSJ/MLLR model.

音声特徴量	LDA-MLLT (40 次元)
音響モデル	MLLR 適応 CSJ コーパス
言語モデル	OGVC 664 発話
発音辞書	OGVC 367 語
適応学習データ	テストデータ

表 3 実験条件-CSJ/BNF モデル

Table 3 Experimental conditions - CSJ/BNF model.

音声特徴量	LDA-MLLT Tandem 特徴量 (42 次元)
音響モデル	CSJ コーパス + Tandem アプローチ
言語モデル	OGVC 664 発話
発音辞書	OGVC 367 語
学習データ	1328 発話
入力次元数	440 次元 (40 次元× 11 フレーム)
HMM 状態数	3481
ネットワーク構造	Input:1024:1024:1024:1024:42:1024:HMM

習にはテストデータに使われない, 残りの男性 1 人, 女性 1 人の音声データを用いる. 今回の実験では主に感情音声の音響的な特徴の差によるモデルへの影響を確認するため, 言語モデル, 発音辞書に関しては OGVC データを用いて作成した. 8 感情×4 段階 (平静, 弱, 中, 強) のテストデータ全てに対して実験を行う. それぞれのモデルの条件は表 1, 表 2, 表 3 に示す.

#### 4.2 実験結果

実験結果を図 2 から図 6 に示す. 図 2 は感情の入っていない平静音声, 図 3,4,5 はそれぞれ感情レベルが 1,2,3 の音声をテストデータとして用いた. 図 6 は 4 段階レベルの全ての平均を示している. 縦軸は WER (Word Error Rate) を表す.

図 2 より, 平静音声の認識に関しては全てのモデルで音響モデルとマッチしているため平静音声の認識ではそれぞれのモデルで大きな差は見られなかった. 図 3 以降の感情を含む音声の認識結果は感情の強度が強くなっていくほど認識精度は低下した. 感情に対しての処理を施していない CSJ モデルは感情に強い影響を受け, 最も認識精度が低い結果になった. 今回は言語モデル, 辞書が小さく音響モデルの性能がそのまま結果に現れている. 特に叫び声の多い SUR や FEA のデータでは特に低い精度になっている. 感

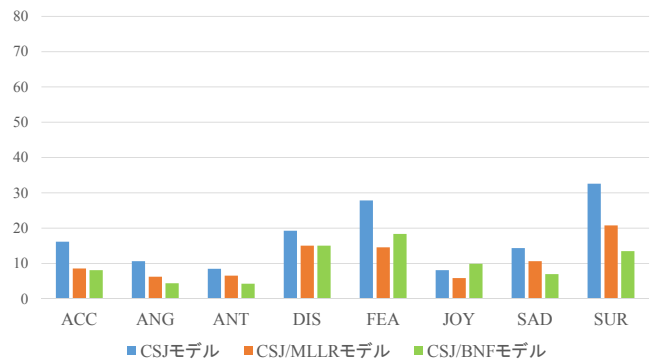


図 2 平静状態音声の認識結果 (WER)

Fig. 2 Neutral speech recognition results (WER) .

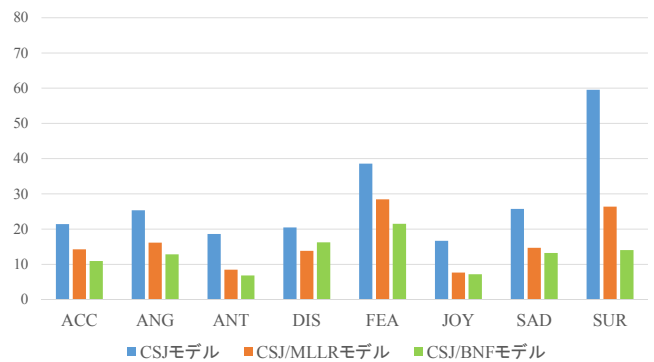


図 3 弱い感情音声の認識結果 (WER)

Fig. 3 Weak emotional speech recognition results (WER) .

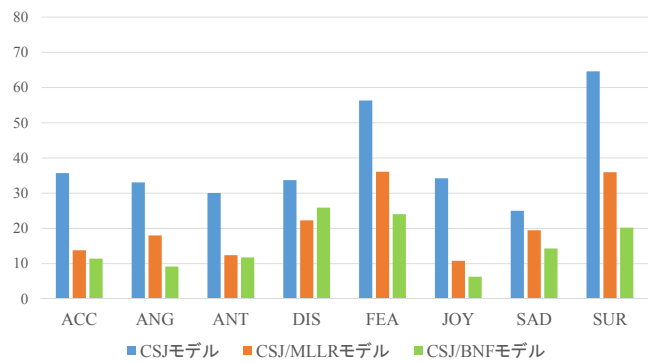


図 4 中程度の感情音声の認識結果 (WER)

Fig. 4 Middle emotional speech recognition results (WER) .

情に対して適応学習を施した, CSJ/MLLR モデルは CSJ モデルと比べて良い結果を認識結果を示した. これは入力音声とフィットした感情を適応データとして用いたためだと考えられる. CSJ/BNF モデルでは多くのテストデータで最も良い認識精度を示した. また認識の難しい叫び声の多いデータも他と比べて良い結果を示している. これは感情の揺れによって判別しにくくなっている音声からでもうまく特徴量を捉えることができたのではないかと考えられる. 感情の強度ごとに平均をとると, CSJ/BNF モデルが最もよく感情音声認識に対して優れたモデルだといえる.

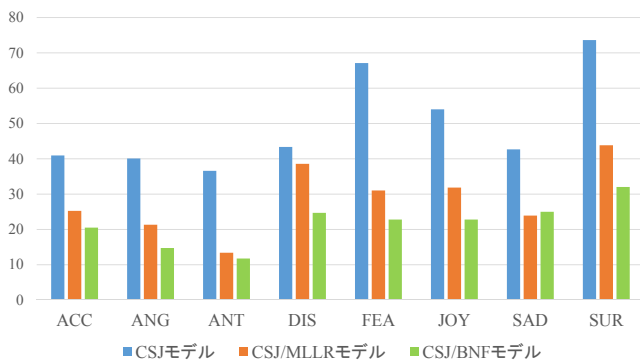


図 5 強い感情音声の認識結果 (WER)

Fig. 5 Strong emotional speech recognition results (WER).

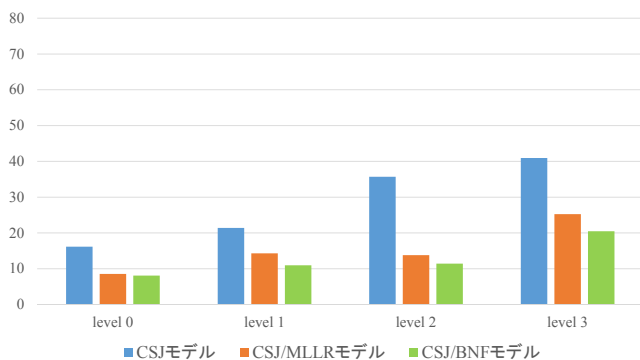


図 6 感情音声のレベルごとの認識結果の平均 (WER)

Fig. 6 Average of each level emotional speech recognition results (WER).

## 5. まとめ

感情音声に対して CSJ モデル, CSJ/MLLR モデル, CSJ/BNF モデルをそれぞれ用いて感情音声の認識を行った。CSJ モデルの結果から感情音声に対して, 特別な処理を施さない場合音声認識結果は著しく低下してしまうことが確認できた。また感情が強くなればなるほど, 認識精度は低下する事を確認できた。感情音声に対する音響的な解決手法として適応モデルの作成は効果を確認できた。

MLLR 適応は話者適応以外にも感情に適応させることで認識結果がよくなるということが言える。しかし今回の感情適応モデルは感情が一致するように作られているため, 複数の異なる感情が入力された場合, 良い結果を示すことができるとは限らない。

音声認識の結果からボトルネック特徴量を用いた CSJ/BNF モデルは感情音声認識に対して有効な手段だといえる。ボトルネック構成のディープニューラルネットワークは 8 種類の感情音声と平静音声を学習データとし, 音素をターゲットとして学習した。その結果, テストでは多くの感情音声, 平静音声の結果が他 2 つのモデルを上回った。音素をターゲットに感情音声の特徴量を入力することで, ボトルネック特徴量には音素の本質的な特徴を

表現できた可能性がある。ボトルネック構成のディープニューラルネットワークに感情音声を学習に使用することで全ての感情音声に対して有効なネットワークを作成することができた。また感情音声の認識精度向上が確認できた。

参考文献

- [1] 篠田浩一, 堀貴明, 堀智織, 篠崎隆宏. 「音声認識」は今後こうなる! 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 2, pp. 1-6, jan 2014.
- [2] 森山剛, 森真也, 小沢慎治. 韻律の部分空間を用いた感情音声合成. 情報処理学会論文誌, Vol. 50, No. 3, pp. 1181-1191, 2009.
- [3] 三村正人, 坂井信輔, 河原達也. ディープオートエンコーダと dnn-hmm を用いた残響下音声認識. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 6, pp. 1-6, jul 2014.
- [4] 門谷信愛希, 阿曾弘具, 鈴木基之, 牧野正三. 音声に含まれる感情の判別に関する検討. 電子情報通信学会技術研究報告. SP, 音声, Vol. 100, No. 522, pp. 43-48, dec 2000.
- [5] Björn Schuller, Jan Stadermann, and Gerhard Rigoll. Affect-robust speech recognition by dynamic emotional adaptation. In *Proc. speech prosody*. Citeseer, 2006.
- [6] Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. Asr for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, Vol. 18, No. 4, pp. 437-444, 2005.
- [7] 波任, 龍標王, 充彦甲斐. ディープニューラルネットワークに基づく音声選択と環境適応による非同期音声収録の音声認識. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 24, pp. 1-6, dec 2014.
- [8] 西崎博光. 音声言語処理のための要素技術と音声ドキュメント処理への応用. 電子情報通信学会技術研究報告. EMM, マルチメディア情報ハイディング・エンリッチメント, Vol. 114, No. 33, pp. 11-16, 2014.
- [9] 渡部晋治. 音声認識における音響モデル (自動音声認識研究の動向と展望). 日本音響学会誌, Vol. 66, No. 1, pp. 18-22, dec 2009.
- [10] 河原達也. 音声認識の方法論に関する考察 世代交代に向けて. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 3, pp. 1-5, 2014.
- [11] 篠田浩一. 確率モデルによる音声認識のための話者適応化技術. 電子情報通信学会論文誌 D, Vol. 87, No. 2, pp. 371-386, 2004.
- [12] 佐古淳, 有木康雄. 知識を用いた音声認識による野球実況中継の構造化 (音声言語応用)(第 6 回音声言語シンポジウム). 電子情報通信学会技術研究報告. SP, 音声, Vol. 104, No. 543, pp. 85-90, dec 2004.
- [13] Hynek Hermansky, Daniel W Ellis, and Shantanu Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, Vol. 3, pp. 1635-1638. IEEE, 2000.
- [14] 久保陽太郎. Deep learning (深層学習)(第 5 回) 音声認識のための深層学習. 人工知能: 人工知能学会誌: journal of the Japanese Society for Artificial Intelligence, Vol. 29, No. 1, pp. 62-71, 2014.
- [15] 福田隆. 音声特徴抽出の基礎と最近の研究動向 (音声・言語・音響教育, 一般). 電子情報通信学会技術研究報告. SP, 音声, Vol. 111, No. 97, pp. 1-6, 2011.
- [16] Jonas Gehring, Yinping Miao, Florian Metze, and Alex Waibel. Extracting deep bottleneck features using stacked auto-encoders. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3377-3381. IEEE, 2013.
- [17] Tang Ba Nhat, 目良和也, 黒澤義明, 竹澤寿幸, Kazuya Mera, Yoshiaki Kurosawa, Toshiyuki Takezawa. 音声に含まれる感情を考慮した自然言語対話システム.
- [18] 有本泰子, 河津宏美, 大野澄雄, 飯田仁. 感情音声のコー

パス構築と音響的特徴の分析 : Mmorpg における音声チャットを利用した対話中に表れた感情の識別. 情報処理学会研究報告. MUS,[音楽情報科学], Vol. 74, pp. 133-138, feb 2008.