Vol.2017-SLP-119 No.7
2017/12/21

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

相澤 佳孝[1]　加藤 正治[1]　小坂 哲夫[1]　能勢 隆[2]

概要：近年，感情音声コーパスとして JTES（Japanese Twitter-based Emotional Speech）が構築された．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．我々はこれまで，日本語話し言葉コーパス (CSJ) を用いて学習した DNN-HMM による音響モデルを初期モデルとして，JTES を用い話者や感情へ適応した音響モデルを用いて認識実験を行なってきた．CSJ で学習し CSJ のテストセットを認識した場合の単語誤り率は 15.12 %と比較的良好な認識結果が得られていた．一方 CSJ による初期モデルを適応した後の JTES による評価を見ると話者適応では 27.86 %と十分な結果が得られていない．本稿では，適応法と言語モデルに関して検討し，更なる性能向上を試みた．適応法としては，DNN のエポック数の決定に early stopping の利用を検討した．また，言語モデルにおいては未知語の影響を調査し，未知語を追加した言語モデルを検討した．以上により話者適応でベースラインが 27.86 %に対し，23.05 %，コーパス適応のベースラインが 32.37 %に対し，26.91 %と大幅な性能向上が得られた．

キーワード：感情音声認識，感情音声データベース，話者・感情・コーパス適応，DNN-HMM

1. はじめに

近年，音声対話システムが注目されている．こうしたシステムは，目的場所への案内や情報検索など特定の目的に用いる場合は機械的なやりとりでも十分実用となる．しかしタスク達成のための対話だけでなく，雑談など対話自体を目的とした用途へと応用が広がりつつある．このような用途の場合，人間同士の会話における感情のやりとりのように，システムが感情を考慮して対話を行うことでより豊かな音声対話が可能になると考えられる．感情を考慮した音声対話の実現のためには，システム側が感情種別を認識でき，感情音声の発話内容を正しく認識し，それらをもとに感情音声を合成できる必要がある．

こうした感情音声を用いた実験に利用できるコーパスがいくつか構築されている [1]．文献 [2] では感情音声認識・合成の観点で音韻・韻律バランスのとれた発話セットの設計を行った．また文献 [3] では感情音声合成の観点で評価し有効性を示している．このコーパスは「JTES」（Japanese Twitter-based Emotional Speech）と名づけられ，今後の応用が期待されている．

文献 [4] では，この JTES を評価データとして音声認識実験を行なった．その際，日本語話し言葉コーパス（Corpus of Spontaneous Japanese：CSJ）[5] のデータをベースとした音響モデルに対して，JTES のデータを fine-tuning により適応し認識していた．しかし，単語誤り率は話者適応を用いても約 28 %であり，十分な結果が得られていない．その原因の一つとして，適応法の検討が不十分であったことが挙げられる．本稿では，DNN の適応時のエポック数に着目し，教師あり適応実験において early stopping を用いた．また，CSJ における学会講演と JTES における Twitter では用いられる語彙も異なり，言語モデルの検討も必要である．このため，未知語や言語モデルの N-gram 確率などを検討する必要があるが，今回は未知語を追加した言語モデルの使用に関して検討を行なった．

2. 感情音声データベース JTES

Twitter の呟きを基にした感情音声データベース「JTES」は，男女各 50 名の音声からなる [3]．Twitter の呟きには口語的なものが多く含まれていることから，様々な感情を付与した発話文を作成することが可能である．Twitter の呟きを感情表現語とのマッチング [6] により喜び・怒り・悲しみ・平常の 4 感情に分類し，音韻・韻律のバランスが

[1] 山形大学
Yamagata University
[2] 東北大学
Tohoku University

取れた文章セットをエントロピーによる文選択アルゴリズム [7] を用いて各感情 50 文の選出が行われる [2]．それらの文に対し主観評価により感情の適切性を確認し，音声の収録が行われた．収録の際には「自分が意図する感情をロボット（機械）に伝えるように」発話し，怒りに関しては「hot anger（激しい怒り）」[8] を意識するように指示している．

3. 適応・認識手法

本章では，本実験で用いられた適応および認識手法について述べる．今回の実験では適応データに正解テキストが準備できることを前提とした教師付き適応を行なう（教師なし適応の結果については文献 [4] を参照のこと）．適応においてはエポック数自動決定のため early stopping を使用する．DNN-HMM を用いた認識においては事後確率補正法を併用する．また言語モデルの未知語対応について述べる．

3.1 early stopping

early stopping は学習や適応時のエポック数を自動決定する手法である．この手法では DNN の fine-tuning を行なう際に，学習（適応）データを開発データと評価データに分割して交差検定を行ない，繰り返しにおいてフレーム認識率の改善率が閾値以下になった場合に繰り返しを停止する．本研究では開発データと評価データの分割割合を 9 : 1 とした．

3.2 事後確率補正

認識時には DNN の事後確率の補正を行う [9]．出力確率計算の際に，無音など一部の音素で状態の生起確率が極端に大きくなる問題に対して対応が可能となる．DNN-HMM における出力確率は式 (1) のように求められる．

$$p(x|s_i) = \frac{p(s_i|x)p(x)}{p(s_i)} \tag{1}$$

このとき，$p(x)$ は入力特徴の生起確率だが認識には影響を与えないので無視する．$p(s_i)$ は状態生起確率だが，無音区間など一部の音素で非常に大きな値となり，出力確率が低下する．従って，この値に制限を加えることで，確率の低下を補償する．具体的には $p(s_i)$ に対して閾値 (上限値)θ を設定し，上限を超えた値を θ に置き換える．この際 θ は式 (2) に基づき決定し，そこで制限率 α $(0 \leq \alpha \leq 1)$ を指定する．

$$\alpha = \frac{\sum_{i \in D}\{p(s_i) - \theta\}}{\sum_{i=1}^{I} p(s_i)} \tag{2}$$

ここで，i は状態番号，I は総状態数を意味し，$p(s_i) > \theta$ を満たす i の集合を D とする．$p(s_i)$ が大きい順に i を並び替えた場合の説明図を図 1 に示す．この方法は適応時など学習 (適応) データが少ない場合特に有効である．

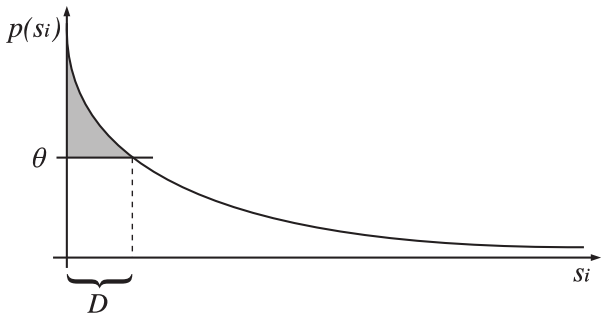

図 1　出力確率補償の説明図

3.3 言語モデルの未知語対応

言語モデルについて，文献 [10] ではタスククローズの言語モデルを用いているが，本実験では音声認識の汎用性を考え，別タスクにより言語モデルの構築を行なっている．学習データ量を確保するために，今回は CSJ を利用した．しかし，CSJ は講演音声であるのに対し，JTES は Twitter の呟きを基にしており，未知語も多く含まれる．そこで今回は単語辞書に未知語を追加することにより，この問題に対処した．

まず，CSJ で出現する単語を基準として，JTES における未知語に関して調査したところ，評価データで 3.15 %の未知語が存在した．このため CSJ から作成した言語モデルを用いた評価の際は，この程度の単語誤りが生ずる．本稿ではこれらの未知語を追加した言語モデルを用いて，未知語の影響を排除した評価を行なう．

4. 実験条件

音声分析では，16 kHz サンプリング，16 bit 量子化の音声信号に対して，分析窓としてハミング窓を用い，フレーム長 25 msec，フレーム周期 8 msec とした．入力特徴は，24 次元のフィルタバンクと対数パワー，及びその 1 次と 2 次の回帰係数を含めた 75 次元で，11 フレームのセグメント特徴として使用する（$25 \times 3 \times 11 =$ 計 825 次元 ）．また，平均分散正規化を行う．

DNN-HMM の構成については，入力層として 825 ノード，隠れ層として 7 層の 2048 ノード，出力層として 3003 ノードとし，出力特徴は HMM の状態確率として用いた．DNN-HMM の学習では，学習ツールキットとして kaldi[11] を用いて，学習データとして CSJ の第 1 刷の学会講演 963 講演を使用した．DNN 学習の諸条件を表 1 に示す．DNN の学習は pre-training では制限付きボルツマンマシン（RBM）を使用する．fine-tuning ではフレーム毎に状態番号ラベルを与え，教師つき学習を確率的勾配降下法（SGD）による誤差逆伝播法で行う．損失関数にはクロスエントロピーを用いる．これらの条件により作成されたモデルを初期モデルとする．

次に言語モデルについて述べる．言語モデルの語彙セッ

表 1　DNN 学習の諸条件

pre-training	
学習係数	第 1 層のみ 0.01〜0.001 へ減少．第 2 層〜第 7 層は 0.4〜0.04 へ減少.
エポック数	10（1 層目のみ 20）
ミニバッチサイズ	1024
モメンタム	0.9（最初の 50 時間データのみ 0.5〜0.9 へ増加）
L2 正則化係数	0.0002
fine-tuning	
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	512

トは CSJ の学会講演及び模擬講演から出現回数 2 回以上の単語を合わせた 49,058 語とする．言語モデルは bigram と trigram の 2 種類を用い，総単語数約 6.68 M の CSJ の学習データより生成する．さらに JTES に出現する未知語を加えた実験を行なうため，JTES のみに出現する単語を Mecab[12] を用いて解析した．解析により得られた 44 種類の単語を単語辞書に追加し，未知語追加モデルを作成した．この場合認識対象単語数は 49102 語となる．

音声認識手法としては，研究室独自の 2 パスデコーダを用いる．この手法では，第 1 パスで triphone と bigram を用いてビームサーチを行い単語グラフを作成し，第 2 パスでは生成した単語グラフを trigram でリスコアし，認識結果を得る構成となっている．音声認識を行う際のビーム幅に関しては単語内 200，単語間 120，仮説数 2400 とした．また，第 1 パスは言語重み 10，挿入ペナルティ −8 とし，第 2 パスは言語重み 12，挿入ペナルティを −24 とした．事後確率補正の制限率は 5 節の予備実験をもとに 0.1 とした．

評価データとして，JTES の中から 400 発話（10 文 × 4 感情 × 10 話者）を用いた．また適応データとして，評価に用いない文を用いた．具体的には，表 2 の通りである．なお early stopping でエポック数を決定する際は適応データを 9:1 に分割し 1 割を交差検定用データとして用いる．各適応の実験条件について以下に示す．いずれも教師つき適応を実施した．

話者適応　同一話者のデータで適応を行った．感情に対しては非依存となる．

コーパス適応　初期モデルの学習に使用した CSJ と評価対象の JTES では音響環境が大きく異なるため，JTES の環境へ適応した．適応データ量が多いという特徴がある．話者・感情非依存の適応となる．

感情適応　特定の感情に適応した感情依存モデルの性能を図るための実験．話者に対しては非依存となる．

話者感情適応　話者および感情の両方へ適応した場合の実験．音響環境としては一番マッチした条件となるが，適応データ数は少ない．話者・感情依存の適応となる．

表 3　DNN 適応の諸条件

学習係数	0.0001
ミニバッチサイズ	2048
モメンタム	0
L2 正則化係数	0.0002
エポック数	交差検定によりフレーム認識率向上が 0.005% 未満の場合停止

表 4　各制限率における認識結果 (WER[%])

制限率	0.00	0.05	0.10	0.15
Average	39.33	37.79	36.12	36.75

適応後モデルを用いた認識実験においては，5 節の予備実験をもとに学習データ（CSJ）から算出した状態生起確率を用いた．また，DNN 適応の諸条件は表 3 のように定めた．本稿では，表 3 におけるエポック数を 5 と固定した場合がベースラインとなる．この条件において，話者感情適応の実験に関しては，適応データ数が 40 しかなく early stopping の交差検定が正しく行えなかったため，ベースラインのみの結果を示す．

5. 予備実験

各種パラメータを決めるため予備実験を行った．いずれも評価データの話者を 2 名に減らした 80 発話（10 文 × 4 感情 × 2 話者）を使用した．事後確率補正の制限率を決めるため，ベースライン条件で種々の制限率を与え，単語誤り率を求めた．結果を表 4 に示す．この結果から制限率 0.1 が最良値であったため，今回のすべての実験で制限率を 0.1 に固定した．また適応時の状態生起確率 $p(s_i)$ について，CSJ の学習データから算出した場合と適応データから算出した場合の，いずれが良いかを調査するため「コーパス適応」の条件で検討した．なお，このときのエポック数は 1 とした．実験より学習データから生起確率を求めた場合の単語誤り率は 24.31 %だったのに対し，適応データから求めた場合の単語誤り率は 79.31 %と大幅に性能が低下した．これは適応データ数が少なく適切に生起確率が求まらなかったためだと考えられる．以上より，CSJ の学習データから算出した状態生起確率を使用する．

また，参考のために初期音響モデル，言語モデルの性能を CSJ の testset1 の学会男性 10 講演と JTES の評価データ 400 文で確認した．CSJ の testset1 では，認識条件の第二パスの値を言語重み 14，挿入ペナルティ −8 とし，事後確率補正を使用しない場合で，単語誤り率は 15.12 %であった．JTES の評価データ 400 文では，認識条件の第二パスの値を言語重み 14，挿入ペナルティ −8 とし，事後確率補正 $\alpha = 0.1$ とした場合で，単語誤り率 38.10 %，音素

表 2　適応実験とそれに対応する適応データ

名称	適応データ数	適応モデル数
話者適応	160 発話（40 文 × 4 感情 × 評価話者 1 話者）	10（ = 評価話者数）
コーパス適応	14400 発話（40 文 × 4 感情 × 90 話者）	1
感情適応	3600 発話（40 文 × 評価感情 1 感情 × 90 話者）	4（ = 感情数）
話者感情適応	40 発話（40 文 × 評価感情 1 感情 × 評価話者 1 話者）	40（ = 評価話者 × 感情数）

表 5　話者適応以外の適応実験の early stopping により決定したエポック数

適応実験	モデル	エポック数
コーパス適応	コーパスモデル	1
感情適応	怒りモデル	2
	喜びモデル	3
	悲しみモデル	1
	平常モデル	2
話者適応		11 〜 23

※話者適応は話者により異なる

誤り率 14.56 %であった．

6. 実験結果

各適応実験の認識結果として単語誤り率（WER）を図 2，音素誤り率（PER）を図 3 に記載する．図において横軸は各適応実験の種別を，凡例については，epoch5 は従来のエポック数 5 で固定した場合，early_stopping は early stopping を使用しエポック数を決定したモデルを使用した場合，estop+unkmodel は early stopping で決定したエポック数の音響モデルの利用に加え，未知語追加モデルを使用した場合を意味する．また，early stopping の際に決定された各適応後モデルのエポック数を表 5 に記載する．

early stopping による教師あり適応実験の結果では，いずれも性能面でエポック数 5 固定時の結果よりも改善していた．最終的なエポック数を確認すると話者適応を除き，1 から 3 程度で停止しており，エポック数 5 固定では全体の適応実験において過学習の傾向があった．

コーパス適応についてエポック数と認識性能の関係を調査した．コーパス適応の結果を図 4 に示す．図 4 よりコーパス適応の場合の最良の認識結果はエポック数 1 で得られているのが分かるが，これは early stopping での自動決定エポック数に一致する．未知語の影響に関しては評価データの未知語率約 3 %に対して，単語誤り率が約 2 ポイント改善するといった結果になった．また，未知語言語モデルを利用し，エポック数を自動決定した場合，話者適応が 23.05 %で最良の結果が得られた．

各適応の効果について考察する．適応前の WER は 5 章の結果より 38.10 %であり，これと比較していずれの適応も効果的であることが分かる．コーパス適応と感情適応を比較すると，WER においてはコーパス適応，PER においては感情適応で若干上回っているが，両者は大差がないと言える．このため今回は感情毎の適応の効果は確認することはできなかった．一方で，感情認識の実験結果として文献 [13] や文献 [14] では，感情強度別で認識することにより感情認識の精度が向上している．これらの結果から，感情強度が異なれば音響的特徴が異なる可能性がある．今後はこの点も考慮して実験する必要がある．また，話者適応はコーパス適応や感情適応よりも高い性能が得られている．このため，話者によるスペクトルの変動は相対的に影響が大きいと言える．

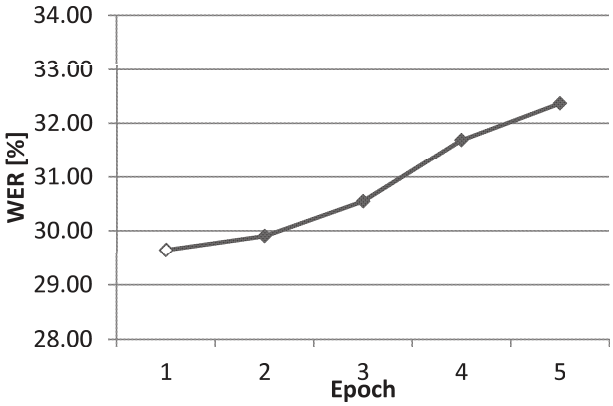

図 4　コーパス適応におけるエポック数 5 までの WER（白抜き部分は early stopping 決定箇所）

7. まとめ

本研究では感情音声コーパス JTES を，適応・評価データとした音声認識実験を行なった．文献 [4] では感情音声データベース JTES を，感情や話者やコーパスといった種別に分割し，音響モデル適応することにより認識精度の向上を試みた．

本稿では更に early stopping の検討と未知語を追加した言語モデルによる評価を行なった．実験結果より，early stopping 使用前の認識精度と比較し，全ての適応モデルの認識精度で改善が得られた．未知語に関してはいずれの適応モデルによる認識でも 2 ポイント程度の改善が得られた．しかし，今回検討した early stopping と未知語を追加した言語モデルによる結果では，コーパス適応と感情適応であまり大差のない性能が得られた．このことから，スペクトルによる感情毎の適応の効果を確認できなかった．この原因として感情表出の強度が話者により大きく異なっていると考えられる．また，early stopping，未知語を追加し

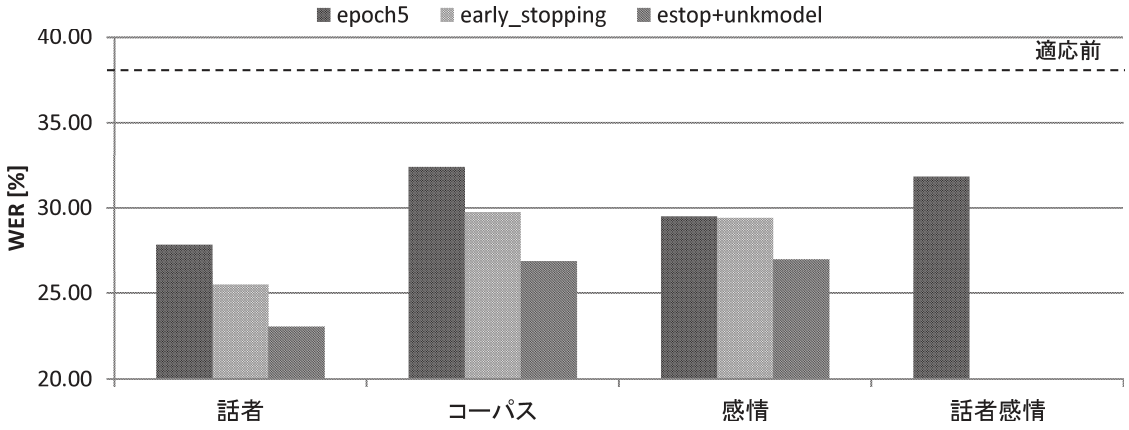

図 2 各適応実験の単語誤り率（話者感情適応については適応データ数が少量のため交差検定が行えず，early stopping 以降の実験を行なっていない）

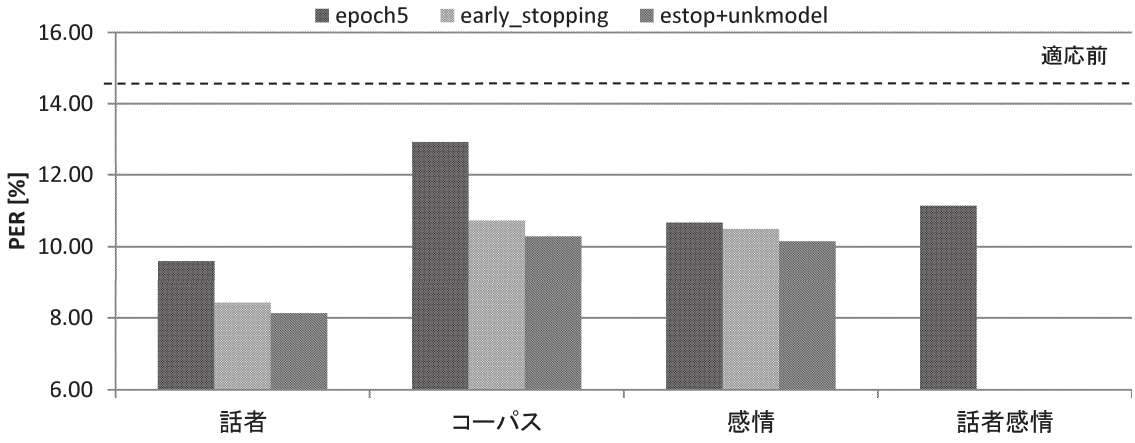

図 3 各適応実験の音素誤り率

た言語モデルを使用した場合のいずれの条件においても，感情音声認識において話者適応が最良であり，23.05 %の単語誤り率を達成した．このことから，話者によるスペクトルの変動は相対的に影響が大きいと言える．

今後の課題として，感情適応の性能向上の検討を行なう．この場合，単純に感情別で適応するのではなく，感情の強度も考慮した適応を検討する．また，今回は CSJ のデータを用いて言語モデルの構築を行なっているが，CSJ と Twitter の発言では大きく単語の出現頻度が異なることが予想される．このため，今後は言語モデルの適応も行ない更なる性能向上を図りたい．

謝辞 本実験の一部は科研費 (課題番号 16K00227, 15H02720）によった．

参考文献

[1] 有本他，“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, 音楽情報科学, Vol.74, 133–138, 2008.

[2] 武石他，“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,”電子情報通信学技術研究報告，115(253)，33–38，2015.

[3] 武石他，“感情音声データベース JTES の音声合成による評価,”音響学会講論集（春），335–338，2017.

[4] 相澤他， “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 DNN-HMM 音響モデル適応の検討,”音響学会講論集（秋），97–100，2017.

[5] K.Maekawa. “Corpus of Spontaneous Japanese: Its design and evaluation,”In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 1–6, 2003.

[6] 高野他，“感情関連語を用いた感情推定法の提案とニュースサイトアクセス解析への応用,”日本感性工学会論文誌,

11，3，495–502，2012.
[7] 荒生他，“対話音声合成のための音韻・韻律コンテキストを考慮した音声コーパス構築法の検討，”音響学会講論集（春），499–500，2013.
[8] Yacoub et al.，“Recognition of emotions in interactive voice response systems，”in *Proc. INTERSPEECH*，729–732，2003.
[9] 冨田他，“高精度な初期モデルを用いた教師なしクロス適応の評価，”音響学会講論集（秋），95–98，2016.
[10] 向原他，“ボトルネック特徴量を用いた感情音声認識の検討，”音響学会講論集（春），43–44，2016.
[11] D.Povey et al.，“The Kaldi Speech Recognition Toolkit，”in *Proc. ASRU*, 2010.
[12] Mecab，http://taku910.github.io/mecab/
[13] 阿部他，“SVM を用いた自発対話音声の感情認識における学習データの検討”，情報処理学会東北支部研究報告，7-A3-3，2016.
[14] 武石他，“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予備的分析，”音響学会講論集（秋），209–212，2016.