Vol.2021-MUS-131 No.32
Vol.2021-SLP-137 No.32
2021/6/18

東京工科大感情語りコーパス (TUT Emotional Story-telling Corpus; TESC) の作成と基礎的分析

大石 光流[1,a)]　榎本 美香[2,b)]　越智 景子[2,c)]　大淵 康成[2,d)]

概要：本研究では，感情的な語りがどのように韻律的な脚色を持つか明らかにするため，東京工科大感情語りコーパス (TUT Emotional Story-telling Corpus; TESC) を作成し，パラ言語情報の基礎的な分析を行う．TESC は，大学生 36 人に自身が経験した嬉しかったことや怒ったことなど 8 つの感情的な語りを行ってもらい，それを収録したものである．OGVC や CSJ の韻律特性と比較し，感情的な対話と独話の違い，感情的な語りと一般的な語りの違いを明らかにする．

Creation and Basic Analysis of the TUT Emotional Story-telling Corpus (TESC)

1. はじめに

語り (storytelling) は [1]，話者が交替する通常の会話とは異なり，一人の話者が自身の体験やある出来事を複数の発話 (Turn constraltional unit;TCU) を用いて語る活動である．語り手は一つの発話を短くゆっくりと話したり，逆に一つの発話を長く早く話したりと，様々な韻律の工夫をし，物語を脚色する．そして，物語のオチとなるクライマックスまで，背景や前景などの情報が繰り返し語られる [2]．

感情的な語りに着目してみると，長田らのうつ病患者の自殺企画前後の感情や状況に関する研究 [3]，栗井らの語りによる感情変化についての研究 [4] がある．しかし，これらは，感情的な語り音声の音響特性を分析したものではない．

一方，感情の音響的特徴を分析したものに，有本ら [5] の音声チャット中の対話に現れた感情の識別に関する分析があるが，話者に自身の感情を表現しようという意図があったわけではなく，対話中に自然に表出された感情であるため，感情の種類を明瞭に識別できるという結果は出ていない．これに対して，声優に感情的な発話を行ってもらった演技音声を用いた感情の識別においては，周波数やパワーといった特徴量において，喜び，悲しみ，信頼，嫌悪，恐れ，怒り，驚き，望みという 8 つの感情で顕著な使い分けがなされていることがわかっている [6]．さらに，一般の大学生に自身の感情的なエピソードを語ってもらった場合には，有本ら [5] の自然音声では特徴の見られなかった恐れ，怒り，望みにおいてパワーの範囲や周波数の最小値において顕著な特性が見られた [7]．これらの先行研究を要約すると，自然発話では感情を表す音響特徴の変化は少ないが，演技音声ではそれが非常に大きくなる．その中間ともいえる素人の感情語りでは，その音響的特徴が半分程度生じる．

本研究ではこの吉田 [7] の研究をより体系的に進めるための第一歩として，東京工科大学感情語りコーパス（TUT Emotional Story-telling Corpus; TESC）を作成する．そして，日本語話し言葉コーパス（CSJ）の一部に含まれる感情的な模擬講演や有本らが公開している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) との比較を通して，TESC の音響的特性を明らかにする．

2. 東京工科大感情語りコーパスの設計と収録

堀内らは，対話コーパスには，(1) 自発的な発話の収録，(2) 心理的・生理的な統制，(3) 音質的な品質の保持といった点が求められると述べている [8]．TESC の収録にあた

1 東京工科大学大学院 バイオ・情報メディア研究科
Graduate School of Bionics, Computer and Media Science, Tokyo University of Technology
2 東京工科大学 メディア学部
School of Media Science, Tokyo University of Technology
a) g312002173@edu.teu.ac.jp
b) menomoto@stf.teu.ac.jp
c) ochikk@stf.teu.ac.jp
d) obuchiysnr@stf.teu.ac.jp

ⓒ 2021 Information Processing Society of Japan

り，今から語ってもらう感情の種類だけを指定し，それ以外は語り手の自発的な創造に任せるようにした．これにより自発的な発話が収録できる．また，語り手が生き生きとドラマティックな表現をできるよう，仲の良い聴き手を防音室越しに対面させ，リアルタイムに合いの手を打つよう設定した．これにより，心理的に日常に近い状態を実現している．また，収録が長時間にならないように 1 人 30 分程度を目処にしており，生理的負担をできるだけ減らしている．防音室の中でヘッドセットマイクを用いて語りをしてもらうことにより音質的な品質も確保している．

2.1 感情の選別

有本ら [5]，池田 [6]，吉田 [7] を踏襲し，プルチックの感情の輪 [9] に見られる 8 感情を取り上げる．「あなたがうれしかったこと・あなたが悲しかったこと・あなたが頼りにしたこと・あなたが嫌悪を感じたこと・あなたが怖かったこと・あなたが怒ったこと・あなたが驚いたこと・あなたが望んだこと」の 8 つである．

2.2 収録手続き

語りを行ったのは，東京工科大学の学部生及び大学院生 18 組 36 名 (男性 33 名・女性 3 名) である．各組の参与者は親近性があり，17 組は同性同士，1 組のみ異性同士のペアであった．

知人である 2 人に収録室へ来てもらい，交互に語り手と聴き手という役割を割り振った．語り手は外部からの音を遮断するために防音室に入り，椅子に座りヘッドセットを装着した．聴き手は防音室とガラス窓を挟んだ外側に座り，ヘッドセットを装着した．

実験者は，聴き手側からガラス越しにこれから話してもらう感情を提示した．用いた感情は，「あなたが嬉しかったこと・あなたが悲しかったこと・あなたが頼りにしたこと・あなたが嫌悪を感じたこと・あなたが怖かったこと・あなたが怖かったこと・あなたが怒ったこと・あなたが驚いたこと・あなたが望んだこと」の 8 つである．語り手ごとにランダムな順番で 1 つずつ提示をし，全 8 種類のお題についての 8 回別々に感情語りを行ってもらった．

これから語る感情を提示し，語り手には 1 分程度で話す内容を考えてもらう．準備ができたという合図が語り手から出されると，聴き手に向かって語りを開始する．

8 つの感情を語り手が語り終わったところで，語り手と聴き手を交代し，再度同様の内容を実施した．両者には謝金が収録時間に応じて支払われる．

収録に用いたヘッドセットマイクは AKG C 420 III であり，そこから得られた互いの音声はハードディスクレコーダー (TASCAM DR-680) により録音するとともに，相手方のヘッドホン (AKG K550) へ分配する．得られた音声は各モノラルチャンネル，48000Hz，24bit の無圧縮の Wav ファイルである．なお参考資料として語り手の上半身をハンディカム (HDR-CX680) にて撮影した．

2.3 発話内容の書き起こし

書き起こしは，語り手のみの音声データを用いた．そのため，聴き手の反応は用いていない．書き起こしの単位として，長い発話単位 (LUU)[10] を用いる．LUU は基本的には文章の句点が入る位置で分割される単位である．しかし，言い淀みや言い差しといった対話独特の言い回しに対応するよう考案されている．会話分析に用いられるターン構成単位 (Turn constructional unit) に近似する．

一つの語り中の一つの LUU にクライマックス (オチ) 記号を付与する．

なお，本報告の時点では，TESC の書き起こしが一部しか完了していないことから，LUU への切り出しができてないコーパス音声の分析は行っていない．書き起こしを行ったのは，被験者 36 名のうち，6 名の 5 つの感情語り（うれしかったこと・悲しかったこと・嫌悪を感じたこと・怒ったこと・驚いたこと）のみである．

3. 分析 1: TESC を用いた音響特徴量による感情の識別

分析 1 では，TESC における各感情語りを識別する音響特徴量を明らかにする．音響特徴量として，基本周波数，音圧，発話速度，モーラ数を用いる．基本周波数と音圧の抽出には Praat[11] を用いる．そして，10 ミリ秒のフレームシフトで抽出し，LUU ごとに基本周波数の平均値 (m_f0) と標準偏差 (s_f0)，音圧の平均値 (m_{int}) と標準偏差 (s_{int}) を算出する．モーラ数 (n_l) は書き起こしテキストをひらがなに変換し，その文字数とする．発話速度 (r) は，モーラ数 (n_l) を発話長 (T) で割ったものである (1).

$$r = \frac{n_m}{T} \tag{1}$$

3.1 一般化線形モデルへのあてはめ

感情ごとに当該感情であるか否かを各音響特徴量から判別する一般化線形モデル (2) へのあてはめを行う．各感情ごとに予測値 (Y_i) と正解率 (p_0) を求める．使用する音響特徴量は，LUU ごとの基本周波数の平均値 (m_{f0})，標準偏差 (s_{f0})，音圧の平均値 (m_{int}) と標準偏差 (s_{int})，発話速度 (r)，モーラ数 (n_l) の 6 つであり，それぞれ x_1 から x_6 である．それぞれの係数 (b_i) は音響特徴がどれだけ予測に影響を与えるかを示す．予測には音響モデルを利用し，閾値を 0.5 とし，閾値以上なら真，閾値未満なら偽の二値分類とする．正解率の算出には (3) を利用した．TP は真陽性，TN は真陰性，FP は偽陽性，FN は偽陰性である．

$$Y_i = b_0 + b_1x_1 + b_2x_2 + ... + b_ix_i \tag{2}$$

$$p_0 = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

語り全体を対象にしたものと語り中のクライマックスとその直前の 10 個の LUU を対象にしたものとをそれぞれ分析する．

3.1.1 結果 1: 語り全体を対象とした識別

結果 1 は語り全体を分析対象とし，一般化線形モデルへのあてはめを行う．分析に用いる音響特徴量はそれぞれ正規化したものを用いる．

算出したモデル式及び正解率は表 1 である．最左端が予測した感情である．例えば，楽しかったことを語っている (1) のか否か (0) を予測するモデルでは，$-1.59+0.11m_{f0}-0.11s_{f0}+0.05m_{int}+0.01s_{int}+0.10T-0.17r^*$ となり，発話速度のみが有意である．なお，右辺冒頭の定数は切片である．各係数の数字の大小はその変数の y への予測の貢献度の大小を表す．正答率は 80%である．

悲しかったことについての感情語りは，基本周波数の平均値 (m_{f0})，音圧の標準偏差 (s_{int}) の係数の値が有意に低く，発話速度 (r) の値が大きい．これは声が低めで一定範囲内に押さえられており，発話速度が速いことを意味する．低音で抑揚が少なく早口であることを示している．

嫌悪を感じたことでは，音圧の平均値 (m_{int}) が大きく，モーラ数 (n_l) がやや少ない傾向にある．声を荒らげて短く話されることがわかる．

怒ったことでは，基本周波数の平均値 (m_{f0}) がやや高く，音圧の平均値 (m_{int}) と発話長 (T) が小さい．少し高めの小声で短く話されるといえる．一般的に怒りは怒鳴り声として大声になると思われるが，語りの中では静かな怒りとして表現されるようである．

驚いたことでは，基本周波数の平均値 (m_{f0}) が小さく，標準偏差 (s_{f0}) が大きい．低めをたどりながら，振れ幅を大きくすることで驚きが表現されている．声を高くすることではなく，低めをつくることで抑揚を出すという脚色がなされているのは興味深い．

3.1.2 結果 2: クライマックスとその直前を対象とした識別

結果 2 は，クライマックスとその直前の 10 個の LUU を対象とし，一般化線形モデルへのあてはめを行う．各音響特徴量は正規化を行っている．

算出したモデル式及び正解率を表 2 に示す．対象データ数が少なくなるため，語り全体を対象にしたもの (表 1) より有意差が見られる項目が少なくなる．裏を返せば，表 2 にみられる有意差は真にその感情の特徴を濃く見せているものといえる．

悲しかったことの発話速度 (r) が速い．嫌悪を感じたことの音圧の平均値 (m_{int}) が大きい．驚いたことの周波数の標準偏差 (s_{f0}) が大きい．悲しかったことは早口で，嫌悪を感じたことは声を荒らげて，驚いたことは音量で抑揚をつけて語られるのは結果 1 と共通して言えることであり，これらの感情を語る上での基本的なテクニックであると考えられる．

4. 分析 2: 既存コーパスとの比較

日本語話し言葉コーパス（CSJ）の一部に含まれる感情的な模擬講演 [12]，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)[13] と TESC の比較を行う．CSJ 中の「人生で一番嬉しかったこと・楽しかったこと」と「人生で一番悲しかった・つらかったこと」をテーマにした模擬講演を用いる．感情的な体験の語りという意味において TESC と共通している．OGVC は自発的な対話中に生じた感情がラベリングされており，さらにそれを声優が演技的に読み上げるという演技音声が収録されている．プルチックの 8 感情が収録されているという点で TESC と共通している．これらのコーパスとの比較を行うことで，講演という完全な独話との違い，自発的な対話中の感情的発話との違い，演技音声との違いを明らかにする．

4.1 分析に用いるデータ

CSJ から上記 2 テーマの模擬講演を用いる．OGVC の 4 名の自発対話 (Vol.1 対話 ID05,06)，2 名の演技音声 (Vol.2,F1,M1) を用いる．

分析に用いる音響特量量は基本周波数と音圧の平均値と標準偏差である．続いて，TESC の LUU の発話長の平均である 2.74 秒ごとに各音声ファイルを切り出した．そして，1 つのファイルごとに基本周波数と音圧を抽出した．抽出方法は TESC の分析時と変わらず，10 ミリ秒で抽出を行い，平均値と標準偏差を取り出した．さらに，無音区間が多い発話区間を除去するため，基本周波数が存在しないものと 2.74 秒のうち，無音区間が 90 %を超えるものを除外した．音圧は sox を用いて最大振幅の音圧の正規化を行った．

4.2 基本周波数の平均値の比較

各コーパスの基本周波数の平均値を箱ひげ図にしたものが図 1 である．CSJ_S01 が「人生で一番嬉しかったこと・楽しかったこと」，S02 が「人生で一番悲しかった・つらかったこと」である．OGVC_vol1 は自発音声，OGVC_vol2 が演技音声を示している．OGVC_vol2 と TESC は，さらに感情別に分かれている．感情はアルファベット順に受容 (ACC)，怒り (ANG)，期待 (ANT)，嫌悪 (DIS)，恐れ (FEA)，喜び (JOY)，悲しみ (SAD)，驚き (SUR) である．

全体的にみると，OGVC の演技音声の値が高め，TESC が低め，その中間に CSJ と OGVC の自発音声がきている．これらの差が録音環境に起因するのは否めない．TESC では静音な防音室での発話なのに対し，CSJ は講演室のため音量を大きくして収録している可能性がある．OVGC も

表 1　語り全体を対象とした音響特徴量のモデル式と正解率

感情語り	音響モデル	正解率
楽しかったこと	$y = -1.59 + 0.11m_{f0} - 0.11s_{f0} + 0.05m_{int} + 0.01s_{int} + 0.10T - 0.17r^{*}$	0.800
悲しかったこと	$y = -1.29 - 0.10m_{f0} - 0.20s^{*}_{f0} - 0.29m^{***}_{int} - 0.18s^{*}_{int} - 0.06T + 0.20r^{**}$	0.794
嫌悪を感じたこと	$y = -1.44 + 0.11m_{f0} + 0.04s_{f0} + 0.45m^{***}_{int} + 0.13s_{int} - 0.14T^{\dagger} - 0.10r$	0.800
怒ったこと	$y = -1.32 + 0.13m^{\dagger}_{f0} - 0.00s_{f0} - 0.15m^{*}_{int} + 0.12s_{int} + 0.16T^{*} + 0.01r$	0.800
驚いたこと	$y = -1.52 - 0.29m^{***}_{f0} + 0.23s^{***}_{f0} - 0.04m_{int} + 0.01s_{int} - 0.07T - 0.11r$	0.794

各定数右肩の † は p<.10，*は p<.05 で，**は p<.01 で，***は p<.001 で有意水準が認められた．

表 2　セット B における音響特徴量のモデル式と正解率

感情語り	音響モデル	正解率
楽しかったこと	$y = -1.41 - 0.16m_{f0} - 0.16s_{f0} - 0.23m_{int} + 0.02s_{int} + 0.03T - 0.21r$	0.829
悲しかったこと	$y = -1.42 - 0.02m_{f0} - 0.36s_{f0} + 0.04m_{int} + 0.03s_{int} + 0.14T + 0.48r^{*}$	0.776
嫌悪を感じたこと	$y = -1.48 + 0.15m_{f0} - 0.23s_{f0} + 0.39m^{*}_{int} + 0.11s_{int} - 0.08T - 0.15r$	0.799
怒ったこと	$y = -1.38 - 0.06m_{f0} + 0.18s_{f0} - 0.20m_{int} - 0.11s_{int} + 0.05T + 0.01r$	0.782
驚いたこと	$y = -1.47 + 0.09m_{f0} + 0.40s^{*}_{f0} + 0.03m_{int} + 0.08s_{int} - 0.08T - 0.31r$	0.816

各定数右肩の † は p<.10，*は p<.05 で，**は p<.01 で，***は p<.001 で有意水準が認められた．

一般的な部屋での収録であり，音量を大きくして収録した可能性がある．そこで，各コーパス内での大小を論じたいと思う．

CSJ では感情間に差が見られない．OGVC の自発音声も CSJ に近い．

一方，OGVC の演技音声では，感情間で違いが見られる．驚き (SUR) の値が最も高く，次いで喜び (JOY)，恐れ (FEA)，期待 (ANT) となる．低い方からみると，嫌悪 (DIS)，怒り (ANG)，驚き (SUR) である．TESC はこの図でみるとほとんど差がない．先の分析 1 の結果から大小を述べるとすれば，怒り (ANG) が高く，悲しみ (SAD)，驚き (SUR) は低かった．悲しみ (SAD) が低いという共通点がある．一方で，演技音声では怒り (ANG) が低いのに対して，TESC では高い．一般的な怒りの表現として「唸り」のような低い声を想像することができる．演技ではそういった表現がなされているのであろう．一方，自発的な語りにおいては，声が高くなる．声を高くすることで怒りのエネルギーを表していると考えられる．また，演技音声では驚き (SUR) が最も高いのに対し，TESC では低くなる．びっくりしたときに奇声をあげるイメージだが OGVC はその通りである．一方，語りでは驚いた瞬間ではなく，それにまつわる出来事をとうとうと話しているので，声は高くならない．

4.3　基本周波数の標準偏差の比較

基本周波数の標準偏差を箱ひげ図にしたものが図 2 である．標準偏差はコーパス間での差はそれほど大きくない．

OGVC の演技音声は感情によって振れ幅が大きいものがある．驚き (SUR)，喜び (JOY)，怒り (ANG)，期待 (ANT) が高い値を示し，分散が大きいことが分かる．TESC はこの図では各感情であまり差がみられない．しかし，分析 1 の結果から驚き (SUR) の標準偏差が大きくなることが分かっている．OGVC も驚き (SUR) の標準偏差が最も大きくなることと一致する．声の声域を広くして抑揚をつけることで，驚きが表現されるということは確かなようだ．

4.4　音圧の平均値の比較

音圧の平均値を図 3 に示す．OGVC の音圧が低くなるのは収録音量の影響によると考えられる．

ここでも OGVC の演技音声の感情による差が大きくなる．悲しみ (SAD)，嫌悪 (DIS) が大きく，驚き (SUR)，怒り (ANG) が小さくなっている．TESC の方は，図から怒り (ANG) が少し小さくなっているようにみえる．こちらも分析 1 とつきあわせると，TESC は嫌悪 (DIS) が大きく，怒り (ANG) は小さい．これは OGVC と同じ傾向である．嫌悪 (DIS) は声を大きくして荒らげ，怒り (ANG) は小さめの声で静かに表現されるといえる．

4.5　音圧の標準偏差の比較

音圧の標準偏差を図 4 に示す．OGVC の自発音声の振れ幅が大きくなっているのは，ゲーム内で敵と接敵するなど，とっさに大きな声を出す場面があるためであると考えられる．CSJ の感情間で差は見られない．

これも OGVC の演技音声に揺れが見られる．標準偏差が大きいものは，驚き (SUR)，喜び (JOY)，期待 (ANT) である．小さいものは嫌悪 (DIS)，怒り (ANG)，悲しみ (SAD) である．TESC の方は図からは差分はよく分からない．分析 1 では悲しみ (SAD) の値が小さくなっており，OGVC とも一致する．悲しみは音量を一定にすることで表現されるといえる．

5.　まとめと今後の課題

TESC の特徴分析 1 と分析 2 の結果をまとめると表 3 のようになる．怒りの基本周波数の平均値 (m_{f0}) が高くなる

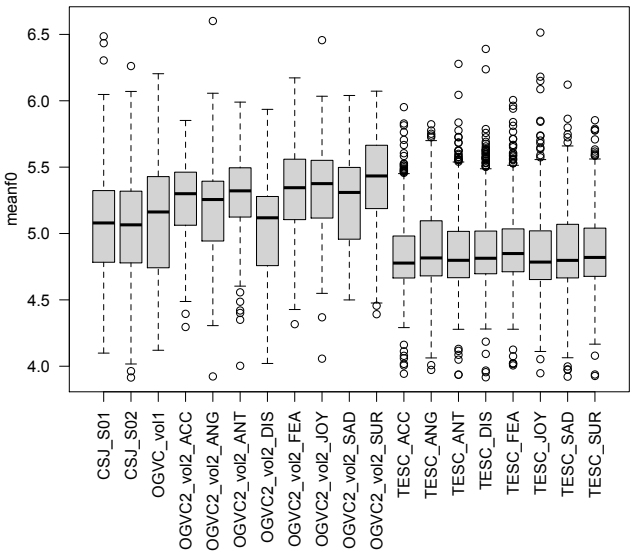

図 1　基本周波数の平均値

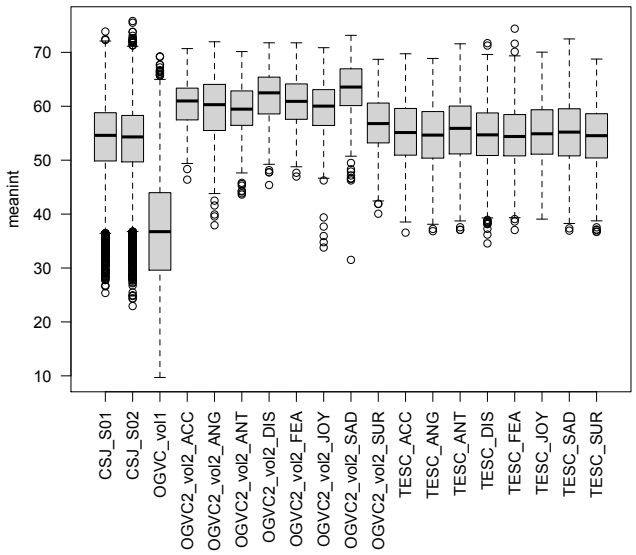

図 3　音圧の平均値

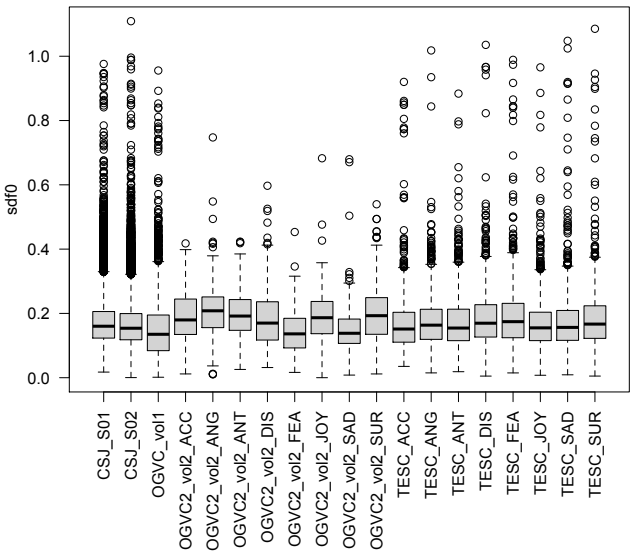

図 2　基本周波数の標準偏差

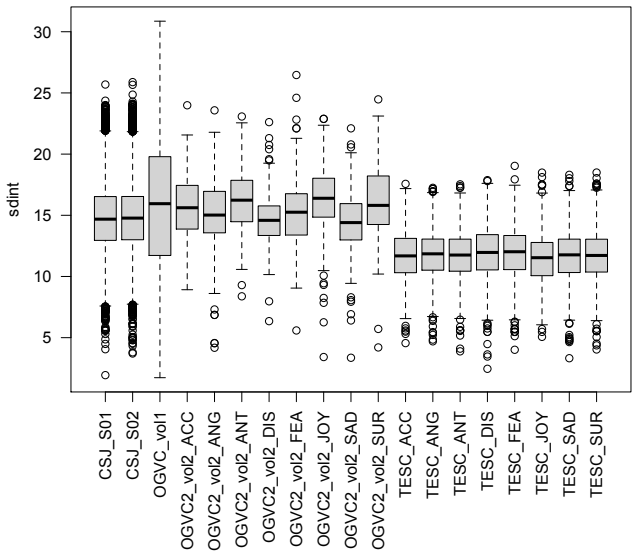

図 4　音圧の標準偏差

ⓒ 2021 Information Processing Society of Japan

表 3　感情ごとの音響的特徴

感情	音響的特徴	語り方
悲しみ	$m_{f0} \uparrow s_{int} \downarrow r \uparrow$	低音で抑揚無く早口
嫌悪	$m_{int} \uparrow T \downarrow$	声を荒げて短く
怒り	$m_{f0} \uparrow m_{int} \downarrow T \downarrow$	高めで静かに長く
驚き	$M_{f0} \downarrow s_{f0} \uparrow$	低めで振れ幅を大きく

のが本コーパスの特徴で，OGVC の演技音声は逆に低くなるのと対象的である．語りの中で怒りは，高い声でそのエネルギーを表現されるのだといえる．その他の特徴量についてはすべて OGVC と共通しており，一般的な感情の表現テクニックであるとみることができる．TESC では演技音声ほど顕著ではないが，少なからず同じ表現が使われていることがわかる．

今後の課題は多い．音圧のレベルがコーパス間で一定になるよう正規化を行う必要がある．また，感情間における個人差を考慮する必要もある．また，TESC の書き起こしが終わっているデータが少なかったため，コーパス間比較においては，時間 2.7 秒窓での分析を行ったが，今後はきちんとすべてのデータにおいて LUU ごとの分析窓にする必要がある．

参考文献

[1] H. Sacks, *An analysis of the course of a joke's telling in conversation*, In J. Sherzer and R. Bauman (eds.), Explorations in the Ethnography of Speaking, Cambridge: Cambridge University Press (1974). pp. 337-353.

[2] Couper-Kuhlen, Elizabeth and Selting, Margret: *Interactional linguistics: Studying language in social interaction*, Cambridge University Press (2017).

[3] 長田恭子, 長谷川雅美: 自殺企図後のうつ病者の企図前・後における感情および状況の分析*:* ナラティヴ・アプローチによる語りから, 日本精神保健看護学会 (2013).

[4] 粟井京子, 内藤直子: 不女性のナラティブ *(*語り*)* による不妊体験の感情変化とビリーフの研究, 香川大学医学部看護学科 (2009).

[5] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida: *Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment*, Acoustical Science and Technology (2003).

[6] 池田裕斗: 技音声に含まれる情報から分析する感情強度, 2012 年度東京工科大学卒業論文 (2012).

[7] 吉田達矢: 感情的エピソードの語りに見られる韻律特性の分析, 2014 年度東京工科大学卒業論文 (2014).

[8] 堀内靖雄，中野有紀子，小磯花絵，石崎雅人，鈴木浩之，岡田美智男，仲真紀子，土屋俊，市川熹: 日本語地図課題対話コーパスの設計と特徴, 一般社団法人 人工知能学会 (1999).

[9] R. Plutchik: *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice*, JSTOR (2001).

[10] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, N. Yoshida: *Two-level Annotation of Utterance-units in Japanese Dialogs: An Empirically Emerged Scheme.*, LREC (2010).

[11] Praat : 入手先 ⟨https://www.fon.hum.uva.nl/praat/⟩（参照 2021-05-17）

[12] K. Maekawa: *Corpus of Spontaneous Japanese: Its design and evaluation*, ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (2003).

[13] 有本泰子, 河津宏美: 音声チャットを利用したオンラインゲーム感情音声コーパス, 日本音響学会 2013 年秋季研究発表会講演論文集 (2013)