多様な話者による演技感情音声の収集と特徴の比較

宮島 崇浩（早稲田大学人間総合研究センター）†
菊池 英明（早稲田大学人間科学学術院）

Collection of Acted-Emotional Speech Using Various Actors and Comparison of Their Acoustical Features

Takahiro Miyajima (Advanced Research Center for Human Sciences, Waseda University)
Hideaki Kikuchi (Faculty of Human Sciences, Waseda University)

１．はじめに

音声研究は、非言語的情報を高度に取り扱うことが求められる段階に至っている。この潮流を受け、近年の音声コーパス研究では表現豊かな音声の収集方法が大きく着目されている(Erickson(2005))。近年は自発音声の収集が主流であるが(Campbell(2005))、演技音声を用いれば、表現豊かな自発音声の収集において制御が難しいとされる録音品質および表現の多様性の確保や、様々なメディア・職業において出現しうる幅広い音声表現の獲得の可能性がある。我々はこれまでに、様々なコンテキストを設定した「台本」を設計し、これに基づいて収録した声優による演技感情音声の音響的・心理的特徴を分析してきた(Miyajima(2012))。そして、提案手法による演技音声は、従来の演技音声と比べて自然性や表現の多様性が向上することが示唆された。本稿では、声優・映像系俳優・舞台系俳優の男女各１名による演技感情音声を比較し、話者・発話内容毎の演技音声の特徴を考察する。

２．多様な音声表現コーパス（SEN コーパス）

我々は、演技者に与える刺激（台本）を工夫することで、表現豊かな音声の収集を試みてきた。この手法によって得た音声資料群を多様な音声表現コーパス（通称：SEN コーパス）と呼称する。図１に収録のコンセプト、表１に工夫した刺激の例を示す。この刺激は、演劇論(安藤(2002))や音声のコミュニケーションモデル(Scherer(2003))を考慮し、構成（図１における『台本のフォーマット』）を決めたものである。この構成に基づき、具体的なインスタンス（内容）を用意し、音声を収録するまでが音声資料生成のスキームとなる。理想的には、より多様な心理パラメータ・音響パラメータ・刺激の内容のバリエーションを確保することで、利便性の高いコーパス構築が可能となると考えた。

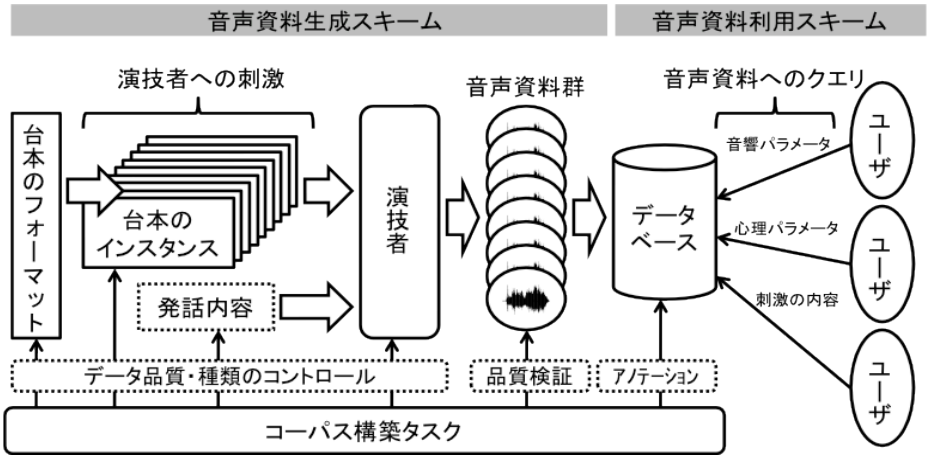

図１　提案手法のコンセプト

†miyajima@toki.waseda.jp

表 1　台本のフォーマットとインスタンスの例

台本のフォーマット		台本のインスタンスの例
共通の情報	発話時の場所・状況	ドラマ、学生寮のホール
	発話者と聞き手の関係	同じ学校、寮に住む親しい友人たち
話し手が持つ聞き手の情報	年齢・性別	17〜18 歳、男
	職業・続柄	高校生
	人物像	騒々しく、騒いでいる最中
話し手自身の情報	年齢・性別	17 歳、女
	職業・続柄	高校生
	人物像	男子校に入った、男のフリをした女 冷静だが、恋愛に関して鈍感
	発声時の背景	周りの騒ぎっぷりを呆れてみるように

表 2　収録済み音声一覧

収録 ID/ フォーマット	台本作成方法/ 発話内容	台本数（音声数）/ 演技者
SEN-1 パターン(1)	TV 番組から主観的に作成（パイロット版） 「ああ、そうですか」	304(100) 女性 1 名
SEN-2 パターン(1)	SEN-1 の項目組み合わせ＋感情表現を明示的に追加 「ああ、そうですか」	280(1400) 女性 5 名
SEN-3 パターン(2)	パーソナリティが変化するよう独自の仕様を作成 「よろしくお願いします」	1000(1000) 男女各 1 名
SEN-4 パターン(3)	SEN-1 の「発話時の背景」を長文化かつ体系化 「いたい」「からい」「いそがしい」など 6 種類	2400(7200) 男女各 3 名
Typical 基本感情語	基本感情語をそのまま刺激として使用 「ああ、そうですか」	80(480) 女性 3 名

2．1　これまでの収録

本手法を用いてこれまでに 3 回の収録を実施してきた（表 2 の SEN-1〜SEN-3)。本稿では，SEN-1 および SEN-2 の概観について触れたのち，最新の収録である SEN-4 の詳細について述べる。SEN-4 は、SEN の拡張として、複数の演技者カテゴリ（声優・舞台系俳優・映像系俳優）および複数の発話（6 種類）を用意している。そこで、それらの違いによる効果の差異を確認する（4 章)。なお、表 2 における Typical は、従来研究で用いられてきたように、基本感情語をキーに発話した感情演技音声の模倣として作成したものである。

2．2　SEN-1 および SEN-2 の特徴

SEN-1 はパイロット版であり、SEN-2 はパイロット版からの話者の拡張および台本インスタンスの改善を施したものである。SEN-1 では、Typical との自然性および心理的・音響的多様性の確認を主に考察し、SEN-2 では話者ごとの効果の違いについて主に考察している。なお、SEN-2 における話者の拡張では、演技者カテゴリを考慮せず、声優経験あるいは舞台経験のある女性 5 名を主観的に選んでおり、SEN-4 ではそれを体系化した点で異なっている（3 章で詳述)。

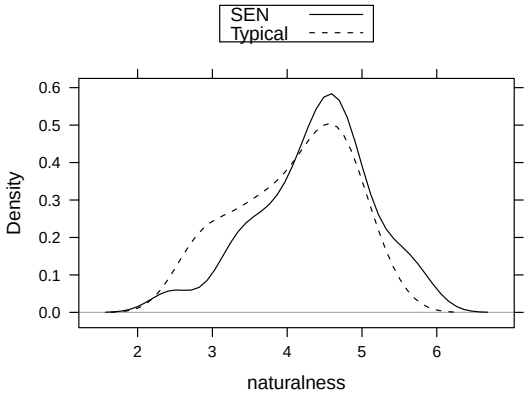

図 2　SEN-1 の自然性評価

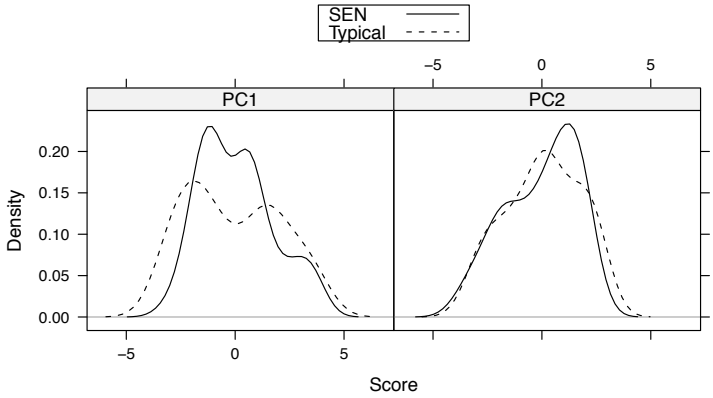

図 3　SEN-1 の心理的特徴の評価（PC1：感情価、PC2：覚醒）

２．２．１　自然性の確認

SEN-1 と Typical の各音声データ群の中から、各 50 個をランダムに抽出し、自然性の印象評定を実施した。評価は、「非常に不自然(1)～どちらでもない(4)～非常に自然(7)」の 7 段階を設定した。図 2 は、評価値に対するカーネル密度プロットである。カーネルは Gaussian カーネル、バンド幅は Silverman の手法により決定した（次節のカーネル密度プロットも同様）。評価値 4 を中心として、自然性の値が低い部分では Typical の密度が全体的に高く、高い部分では SEN-1 の密度が全体的に高いことが分かる。また、各データ群に対して t 検定を実施したところ、p=0.06 で有意傾向であることを確認した。以上より、本手法によって収集する音声は Typical と比べると自然性が高くなる傾向があることが読み取れる。

２．２．２　心理的特徴の確認

自然性同様、SEN-1 と Typical 各 50 個のデータに対して心理的特徴の測定を試みた。心理的特徴の測定方法として、森山(1999)による「音声による感情表現語」を評価語とした印象評定を実施した。感情表現語は、「怒り・喜び・皮肉・恐れ・悲しい・驚き・こび・穏やか・おかしい」の 9 語からなる。それぞれの評価語に対して、「全く当てはまらない(1)~どちらでもない(4)~非常に当てはまる(7)」の 7 段階を設定した。評価者は、大学生の男女各 6 名を選んだ。図 3 は、評価結果に対して主成分分析を実施し、第一主成分(PC1)と第二主成分(PC2)のカーネル密度プロットを表示したものである（第二主成分までの累積寄与率 73%）。第一主成分は感情価（valence）、第二主成分は覚醒(arousal)と解釈した。覚醒の観点では SEN-1 と Typical はほぼ同様であるが、感情価の観点ではいわば相互補完の関係にあることが分かる。SEN-1 は感情価の強度が強い音声表現が少ないため（明確に感情表現を指定することをできるだけ避けたため）、SEN-2 では感情表現を加えた台本を設定して音声を収録した。

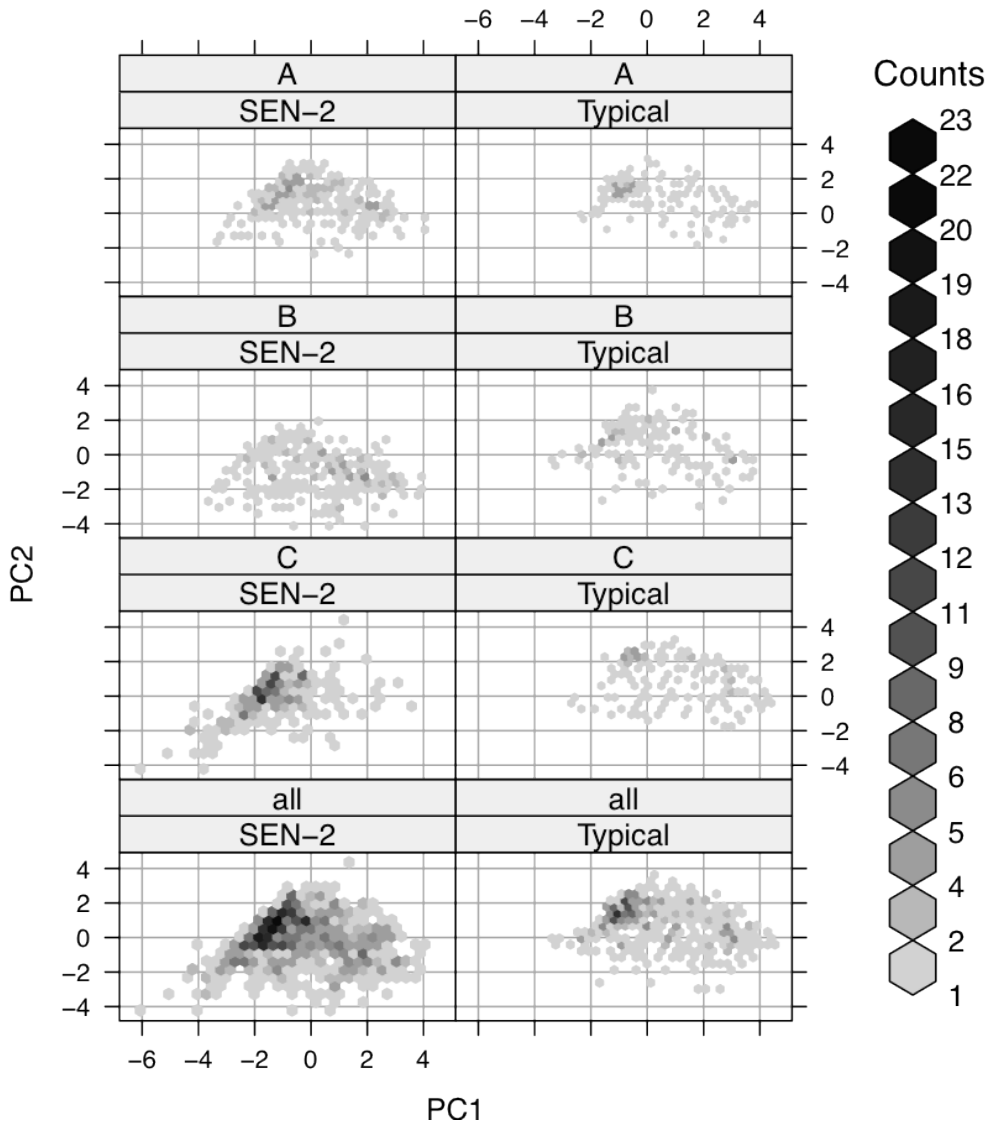

図 4　SEN-2 の音響的特徴の概観（女性 3 名分）

２．２．３　音響的特徴の確認

図 4 は、SEN-2（1 話者につき 280 個）と Typical（1 話者につき 160 個）の音響的特徴を比較したプロット図である。このプロット図では、Carr(2011)の手法により、散布図は大きさが変化する六角形の色の濃度によって密度が表現されている。音響的特徴として、一発話全体に対する F0 関係 5 種（平均、最大、最小、レンジ、標準偏差）および発話長関係 3 種（発話長、平均モーラ長、合計ポーズ長）を選び、主成分分析を行った。累積寄与率は、第二主成分までで 61%である（ここでは、概観を確認するため第二主成分までを示す）。第一主成分は発話内の F0 の変動、第二主成分は F0 の高さや発話長の成分が強く現れている。発話者は、SEN-2 と Typical の両方で収録を行った女性話者 3 名である。Typical は話者ごとの違いがほとんど見られないのに対し、SEN-2 は話者ごとによる違いが大きいことが分かる。つまり、SEN で採用した、演技に必要な多様な情報を刺激として提示するという手法は、話者ごとに表現の拡大の傾向が異なるという結果を導くことになり、これは先に述べた演劇論(安藤(2002))における、熟達者はひとつの脚本解釈から多様な演技計画を立てるという考察と合致していると考えられる。また、3 名分全体(all)で見ると、SEN-2 の音響的特徴は Typical から範囲がほぼ相似形を保ったまま拡大されていることが分かる。

３．多様な発話者カテゴリ・発話内容による SEN コーパスの拡張（SEN-4）

３．１　SEN-4 の収録の目的

SEN-1 および SEN-2 の心理的・音響的分析によって、本手法の適用による自然性の向上および心理的・音響的多様性向上の可能性が示唆された。とくに、２．２．３節で示した、SEN-2 の音響的特徴において話者ごとの違いが顕著に確認されたことは､本研究における要点だと考えられる。我々は、次の課題として、演技者のカテゴリや発話内容の違いによる多様性の広がりの差異について調べることにした。この目的のため、SEN−4 では、発話者（演技者）および発話内容を精密に選定することにし、さらに、表現豊かな発話を促すため、台本の改善も試みた。

３．２　発話者の選定

３．２．１　発話者カテゴリの考察

これまでの収録では、声優や舞台俳優としてのキャリアを持つ人物を主観的に選定してきた。しかし、SEN-3 までの収録した音声を聴取し考察するなかで、同じ声による演技のプロフェッショナルでも表現方法の違いはかなり大きく、本手法における音声表現の多様性に大きく左右することが分かってきた。そこで、いくつかの仮説を考えた。

> 仮説１：声のみによる表現と複数モダリティによる表現では表現方法が異なる
> 仮説２：聴衆者をどのように意識して演技するかで表現方法が異なる

これらの仮説に基づいて､我々は声を用いたプロフェッショナルの演技者のカテゴリ（発話者カテゴリ）として、次の区分を定義した：

- 声優：声のみを用いた表現者、観客は意識しない
- 舞台系俳優：複数モダリティを用いた表現者、観客を意識する
- 映像系俳優：複数モダリティを用いた表現者、観客を意識しない

さらに、3 カテゴリに関して、次の仮説を考えた。

> 仮説３：声の表現の多様性は、映像系俳優、舞台系俳優、声優の順に高くなる
> 仮説４：声の表現の自然性は、声優、舞台系俳優、映像系俳優の順に高くなる

本稿ではまず、仮説 3 における音響的多様性の確認をおこなう。声優は、声のみで多様な表現を求められるため、最も多様性が高いと想像できる。また、舞台系俳優と映像系俳優は、観客に対して演技することを意識するかどうかで、表現の度合いの大きさに差が出る可能性がある。例えば、舞台系俳優は観客を意識した演技をすることに対して、映像系俳優は観客を意識せず、日常生活に近い演技を行っていると考えられる。なお、観客の存在を意識して演技する点の重要性は、先行研究(安藤(2002))で述べられている。

以上のように、キャリアに基づく演技手法の質の違いにより、本手法による同一の手続きに従った演技でも、表現の広がりの差異を生み出すことが期待できる。また、その差異は、本研究の目的である、表現の多様性の獲得に強く関連するであろう。

さらに、男女による表現の違いも検討する。これまでの収録では主に女性を中心に収録を実施してきたが、基本周波数の特性の違いを考慮すると、女性の音響的な表現力が豊かである可能性が高い。これを検証するため、SEN-4 では、各発話者カテゴリについて、男女を別カテゴリとして扱い、計 6 カテゴリの話者を選定することにした。

表 3　SEN-4 における発話内容の詳細

セット	発話内容	極性	音声単語親密度	アクセント型	末尾の音素列
セット１	うまい	Positive	6.312	2	/ai/
３モーラ	からい	Neutral	6.250		
形容詞	いたい	Negative	6.312		
セット２	すばらしい	Positive	6.344	4	/asii/
５モーラ	いそがしい	Neutral	6.156		
形容詞	むずかしい	Negative	6.031		

３．２．２　オーディションの実施

演技者の選定は、オーディション形式を採用することにした。手続きとして、12 通りの台本サンプルを用意し、発話を録音してもらったうえ返送してもらうよう依頼した。台本サンプルの形式は、SEN-4 で用いた台本（表 1 の「発話時の背景」を長文化したもの）と同一であり、発話内容は「ああ」と「そうですか」の 2 種類（各 6 通り）を用意した。「ああ」を選んだ理由は、極めて短い発話でも表現力が豊かであるかどうかを確認するためであり、「そうですか」を選んだ理由は、既存の SEN のデータと比較しながら多様性を演出できる人物かどうか検討するためである。

依頼に際しては、声優・舞台系俳優・映像系俳優というカテゴライズを良く理解でき、かつ各カテゴリに対して人脈を持つ人物を仲介人として起用した。仲介人経由で、声優は特定の事務所、舞台系俳優は特定の劇団、映像系俳優は特定の映画監督に依頼し、本研究の主旨を伝えたうえで、適正があると判断したもらった人物を各複数名選定してもらった。

最終的には、サンプル音声を聴き比べ、筆者らの主観で表現力の多様性が同程度であると思われる人物を、各カテゴリから男女各一名選定した。

３．３　発話内容のデザイン

発話内容は、SEN-3 までは中立的な内容であることを条件としてきた（「ああ、そうですか」「よろしくお願いします」）が、２．２．２節で述べた SEN-1 の心理的特徴の傾向の考察から、提案手法は感情価の強度の意識的なコントロールが重要であると考えられるため、発話内容の（感情価の）極性の違いが多様性に及ぼす影響の検証、さらに発話内容の極性毎による音響パラメータと心理パラメータの差異の検証が必要であると考えた。

また、発話の長さによっても多様性に違いが生まれる可能性がある。理想としては、どのような長さでも同様の多様性が確保できることが望ましいが、これまでの収録の過程で、複数の演技者より、短すぎる発話はバリエーションのある演技は難しいという意見が得られているため、これも合わせて検証できるような発話内容を選択することにした。

これらの観点に従って、発話内容の極性および長さを複数用意することにした。極性は、先行研究(小林(2005), 東山(2008))で提案された極性評価が記された辞書に従い、Positive/Negative/Neutral をそれぞれ選ぶことにした。また、発話の長さに関しては、モーラ数およびアクセント型を揃えることにした。さらに、全ての発話において、BPM などの重要な韻律情報の比較ができるよう、末尾の音韻が統一されるようにした。その他、音声単語親密度(天野(1999))や品詞の統一なども考慮した。

以上の条件に従い、図 4 に示した 6 種類の発話を収録することにした。

表 4　変更前後の「発話時の背景」の例

SEN-1, SEN-2	まだ見ぬヒーローをカッコいいと妄想し、憧れるように
SEN-4	学校から帰る電車の中。A は、手帳とにらめっこしながら今後数週間は一日も空かずに予定が埋まっていることを改めて確認する。それを見た瞬間、先を考えると過酷・・・と思いながら出た一言。

表 5　インスタンスに付与した属性一覧

属性	内容	個数
話し手の年代	10, 20, 30–40, 50–60 代	各 25 通り（計 100 通り）
聞き手の年代	上記＋なし（独り言）	話し手と同じ年代の場合：10 通り 異なる場合：4 通り、独り言：3 通り
聞き手の性別	同姓、異性	聞き手の各年代で半数ずつ（5 通り、2 通り）

３．４　台本の改善

３．４．１　日常性と非日常性

SEN-1 では、台本のフォーマットに従い、インスタンスを TV 番組からランダムに抽出するという手法を取った。また、SEN-2 では SEN-1 から抽出した項目の組み合わせで台本のインスタンスを拡張した。

SEN-1 および SEN-2 での問題点のひとつとして、Typical に比べて自然性は向上したと考えられるものの、やはり非現実的な場面、すなわちアニメーション等の創作物やバラエティ番組で聴くことができるような、やや偏った表現が多いと思われる点であった。我々の目的は、そのような非現実的な場面での表現も、我々が日常生活下で聴取できるような表現も全て収集できる手続きを確立することにある。

そこで、台本のインスタンス生成の際、意識的に「日常的な」インスタンス、「非日常的な」インスタンスの両方を生成することにした。具体的には、一つの発話につき、日常的な場面を意識した台本を 100 本、非日常的な場面を意識した台本を 100 本用意することにした。非日常的な場面とは、SEN-1 で用いたような、テレビ番組でのワンシーンの再現であり、日常的な場面とは我々が日常生活下で遭遇するワンシーンの再現である。これらを今回は完全に創作で作成した。また、日常性や非日常性をより強調するため、次節で示すように発話時の背景の長文化を施した。台本の原案は発話内容 1 種につき各 200 本の合計 1200 本であるが、発話者の性別による違和感の無いよう改変し合計 2400 本を用意した。

３．４．２　台本フォーマット「発話時の背景」の長文化

SEN-4 の台本フォーマットは SEN-1 と同様であるが、発話時の背景を長文化し、ショートストーリー仕立てにした。これは、これまでの収録で、発話時の背景の情報が少ないために強引に表現を広げようとし、不自然な音声表現を招く可能性があるという指摘を受けてのことである。変更前後の文章の例を表 4 に示す。長文化することで、より自然な表現、また、発話者カテゴリ毎に異なる多様な表現を導くことが目的である。

３．４．３　インスタンスの各項目への属性の事前付与

SEN-4 では、インスタンス作成は完全ランダムではなく、日常・非日常の各 100 本について、話し手の年代・聞き手の年代および性別の組み合わせ（属性）を創作前に指定し、音声データへのクエリとして利用できるようにした。詳細を表 5 に示す。

３．５　収録

国立国語研究所内のマルチメディアスタジオに、パソコンおよびUSBオーディオデバイス(Roland UA- 25)を設置した。マイクはヘッドセットタイプ(SHURE WH20XMR)を採用した。サンプリング周波数は44,100Hz、ビットレートは16bitに設定した。また、バックアップとしてリニアPCMレコーダ(SONY PCM-M10)による録音を同時に行った。さらに、将来的な検証のため、デジタルビデオカメラによる撮影を行った。

収録は、約1ヶ月にわたり断続的に実施した。1名あたり1200発話の収録を4回に分け、1回あたりの時間は発話者の任意で休憩を挟みながら3時間程度であった。刺激の提示はスタジオ内に設置したディスプレイを用いて行った。提示順は全員ランダムであったが、日常的な台本から始めたほうが演技しやすいとのコメントを受け、そのようにコントロールを行った。以上の手続きで、合計7200音声の収録を完了した。

４．収録した音声の音響的特徴の概要

SEN-4の7200個の音声について、２．２．３節で用いた音響的特徴を算出し、主成分分析を試みた。第一主成分はF0関連、第二主成分は発話長関連の成分が強く含まれている。第二主成分までで累積寄与率は72%であった。図5に、3モーラ、5モーラの単語ごと、極性ごと、そして話者ごとの計36通りの音響特徴量の密度プロット（各200個）を示す。

４．１　発話長ごとの分布の傾向

全体の傾向を確認する限り、3モーラ発話と比較して5モーラ発話は分布が狭い傾向が見られた。仮説では、5モーラ発話のほうが表現が豊かになると考えていたが、今回の主成分分析で用いた特徴量で判断する限りでは、3モーラ発話のほうが表現が豊かであるような結果が得られた。これらの特徴量は、発話全体に対して計算したもっとも単純な算出方法で求めたものであったので、短い発話ではそのような単純な特徴量で表現の多様性が演出された可能性がある。モーラ長による表現力の違いをより精密に検証するためには、BPMなど局所的な音響特徴量の確認が必要である。

４．２　極性ごとの分布の傾向

すべての話者に関して、極性ごとに大きな差異は見られなかった。本手法が必ずしもこの傾向を必ず保証することではないが、少なくとも今回用意した台本では、極性に左右されず、多様な表現を生み出すことが出来たと考えられる。

４．３　話者カテゴリごとおよび性別の違いによる分布の傾向

発話者カテゴリおよび性別による表現力の比較であるが、これも仮説と反し、声優と舞台系俳優については男性のほうが音響的多様性が高いことが確認された。特に女性声優については、収録中の表現の主観的印象は非常に幅広かったので、筆者らにとって意外な結果であった。一因として、この分析では用いていない音声パワーを良くコントロールして多様な表現を再現していたので、このような結果になったのだと推測される。

４．４　考察

今回起用した6名の話者が、各発話者カテゴリを代表する特徴を持つとは限らないが、全員が発話内容6種類のいずれもほぼ同様の分布を示したことは興味深い結果である。少なくとも、音響的多様性において、発話内容による影響は比較的小さいと考えられる。

また、男性映像俳優（MAM）の分布がいずれの発話においても一番小さい分布を示しているが、台本に対する整合性は他の話者カテゴリと同程度だと思われた。映像俳優は、非常に微妙な声の表現の際で台本の状況を再現しており、興味深いデータが得られている。

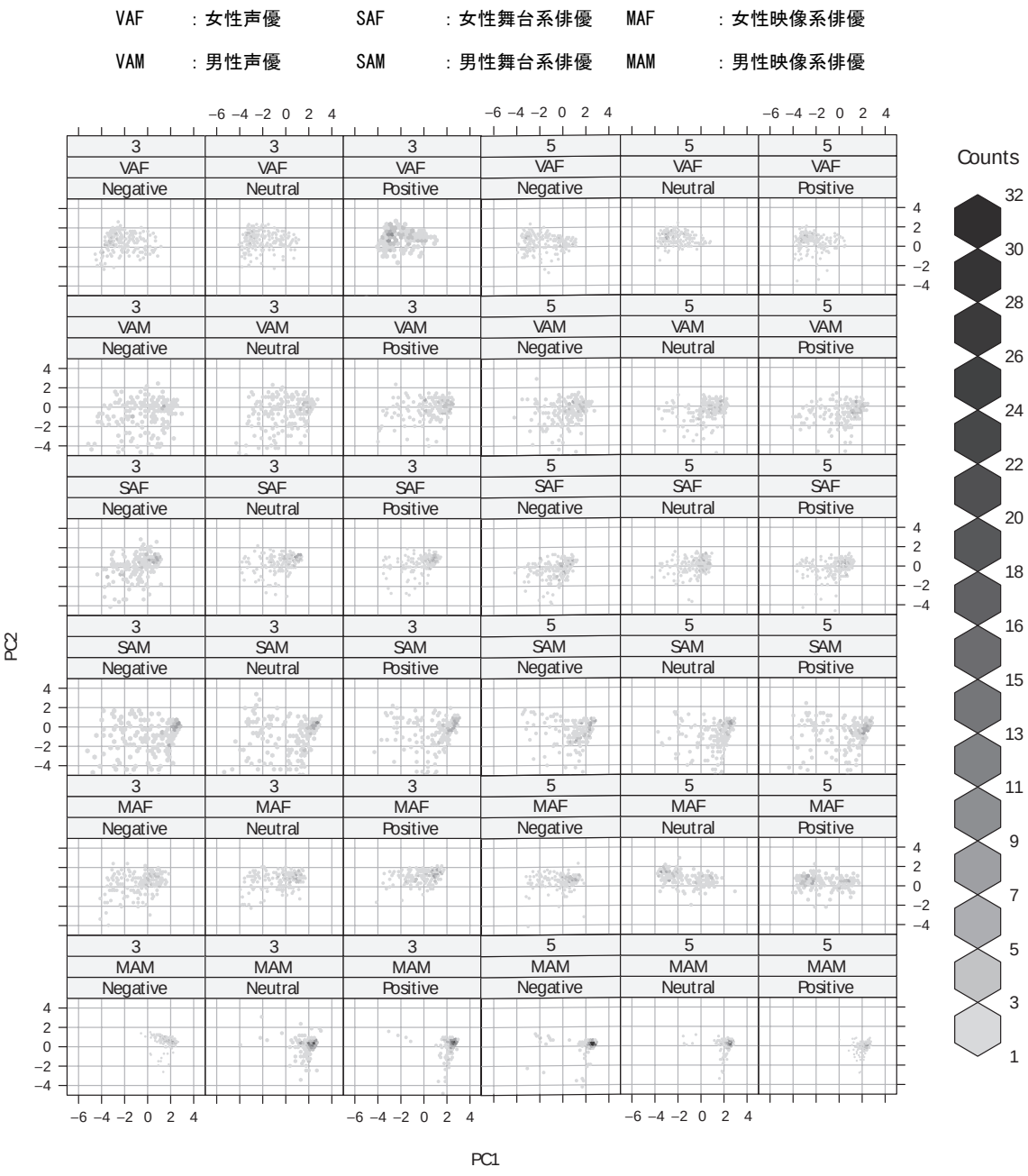

図 5　SEN-4 の音響的特徴の第一、第二主成分の密度プロット（各 200 個）

５．まとめ

我々が提案してきた多様な音声表現コーパスの構築手法に従い、多様な話者カテゴリ・多様な発話内容の収集を大規模に実施した。そして、各発話に対するシンプルな音響特徴量を用いて音声表現の分布の比較を行ったところ、いずれの話者カテゴリにおいても、発話内容の極性が分布に影響していないことが確認された。また、モーラ長が 5 モーラよりも 3 モーラのほうが広い分布を示した。これらの傾向は、選択した音響特徴の影響を強く受けていると考えられる。また、話者カテゴリ毎の特徴に関しては、我々の仮説と異なる結果が得られた。いずれに関しても、多様な心理的・音響的特徴による比較を通した追加検討が必要である。また、各演技者カテゴリの人数を増やすことで、結果がより明確になることが期待できるほか、本稿中で示した複数の仮説の検証が可能となるだろう。

謝　辞

SEN-4 の収録は国立国語研究所(言語資源研究系)基幹型共同研究「コーパス日本語学の創成」(リーダー:前川喜久雄)による成果である。また、演技者の選定や収録に関して、アーティスト(美術家・演出家)の河村美雪氏に多大なご協力を頂いた。最後に、台本に関して、早稲田大学人間科学部の菊池梨佳子氏が 1200 本もの原案を創作してくれた。ここに記して感謝の意を表する。

文　献

D. Erickson (2005). “Expressive speech: Production, Perception and Application to Speech Synthesis” Acoust. Sci. & Tech., vol.4, no.26, pp.317-325.

N. Campbell (2005). “Developments in corpus-based speech synthesis: approaching natural conversational speech” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.376-383.

T. Miyajima, H. Kikuchi, K. Shirai, and S. Okawa (2012). “Method for Collection of Acted Speech Using Various Situation Scripts” Proc. of LREC 2012, pp.1179-1182.

安藤花恵(2002)「演技の熟達化脚本の読み取りから演技計画、演技遂行まで」心理学研究、Vol.74、No.7、pp.373-379.

K.R. Scherer (2003) “Vocal communication of emotion:a review of research paradigms” Speech Communication, vol.40, pp.227-256.

森山剛、斎藤英雄、小沢慎治 (1999) 「音声における感情表現語と感情表現パラメータの対応付け」信学論, vol.J82-D-2、no.4、pp.703-711.

D. Carr, N. Lewin-Koh, and M. Maechler (2011) “Hexbin: hexagonal Binning Routines” R package version 1.26.0.

小林のぞみ、乾健太郎、松本裕治、他 (2005)「意見抽出のための評価表現の収集」自然言語処理、Vol.12、No.2、pp.203-222.

東山昌彦、乾健太郎、松本裕治 (2008)「述語の選択選好性に着目した名詞評価極性の獲得」言語処理学会第 14 回年次大会論文集、pp.584-587.

天野成昭，他(編・著) (1999) 『日本語の語彙特性』、三省堂.