ISSN 2758-2744

音声コミュニケーション研究会資料

Proceedings of the Technical Committee on Speech Communication
The Acoustical Society of Japan

Vol. 3, No. 2
SC – 2023 – 7 ～ SC – 2023 – 13

2023 年 2 月 24 日
February 24, 2023

一般社団法人 日本音響学会

音声コミュニケーション研究会資料目次

Contents

2023年3月3日, 日本音響学会音声コミュニケーション研究会資料, Vol. 3, No.2, SC-2023-7
2023年3月3日, 日本音響学会音声研究会資料, Vol. 3, No.2, SP-2023-10

鼻副鼻腔の音響特性の計算と計測による検証の試み

福島 侑希[†‡] 田島 基陽[†] 竹本 浩典[†⁑]

† 千葉工業大学 〒275-0016 千葉県習志野市津田沼 2-17-1

E-mail: ‡ s19c3105gz@s.chibakoudai.jp, ⁑ hironori.takemoto@p.chibakoudai.jp

あらまし CT データから抽出した鼻副鼻腔の幾何学的モデルの音響特性は，時間領域差分法により計算することができる．そしてその結果は，同じ CT データから造形した声道の実体模型の音響特性を計測することで検証できる．ところがこの計測は成功していない．これは，狭い鼻孔から入力される測定用の音響信号が，声門では十分な信号対雑音比で観測されないためである．この問題を解決するために，本研究では振幅が大きい計測信号を出力可能なエクスポーネンシャルホーンと，壁振動を抑制するために壁を厚くした声道の実体模型を導入した．その結果，伝達関数を計測することに成功し，計算した伝達関数を評価することができた．評価の結果，声道の実体模型では副鼻腔の微細な構造を十分な精度で再現できていないことが示唆された．

キーワード 鼻腔，副鼻腔，声道実体模型，音響計測，FDTD 法

Simulations of acoustic properties for the nasal and paranasal cavities and attempts of validation by acoustic measurements

Yuki Fukushima[†‡], Motoharu Tajima[†], and Hironori Takemoto[†⁑]

† Chiba Institute of Technology 2-17-1 Tsudanuma, Narashino, Chiba, 275-0016 Japan

E-mail: ‡ s19c3105gz@s.chibakoudai.jp, ⁑ hironori.takemoto@p.chibakoudai.jp

Abstract The finite-difference time-domain method can calculate acoustic properties of the geometrical model of the nasal and paranasal cavities extracted from CT data. The calculation results are validated by acoustic measurements of the physical model constructed from the same CT data. The measurements, however, have been unsuccessful. This is because measurement signals input through the nostrils are not observed at the glottis with a sufficient signal-to-noise ratio. To overcome this problem, an exponential horn which can supply measurement signals with high amplitude and a physical model with thick wall to depress the wall vibration were introduced in the present study. As a result, a transfer function was successfully measured, to evaluate the calculated one. The evaluation implied that fine structure of the paranasal cavities could not be reproduced with sufficient accuracy in the physical model.

Keywords Nasal cavities, paranasal cavities, vocal tract physical model, acoustic measurement, FDTD method

1. はじめに

鼻副鼻腔の形状は非常に複雑である．鼻腔は後鼻孔から外鼻孔に至る空間で，鼻中隔で左右に二分され，鼻甲介で緩やかに上下に三分されている．副鼻腔は鼻腔の周囲の硬組織の内部に存在する盲嚢で，前頭洞，上顎洞，篩骨洞，蝶形骨洞からなり，自然口と呼ばれる狭い孔で鼻腔と連絡する（図１）．

鼻副鼻腔の形状は個人差が大きく，音声の個人性の生成要因の一つである[1]．鼻副鼻腔の病変などによる手術で形状が変化すると，音声の個人性が変化することがある．そのため，東京慈恵会医科大学から鼻副鼻腔手術による音声の変化を術前に予測したいという要望が出された．そこでわれわれは，術後の音声を計算機シミュレーションで予測する研究を行なってきた[2, 3]．その結果，術前の CT 画像から 3 次元再構築した鼻副鼻腔の形状データを計算機上で模擬手術し，その形状から音響特性を時間領域差分法（FDTD 法 : Finite-Difference Time-Domain method）で計算することで，術後の音声スペクトルと類似した伝達関数が得られることが明らかになった[4].

しかし，根本的な問題として，CT 画像から 3 次元再構築した鼻副鼻腔の形状データから FDTD 法で計算した音響特性がどの程度の精度を持つのか明らかになっていない．一般に，声道形状から計算した音響特性は，同じ声道形状を光造形などで実体化した声道模型の音響特性を計測することで検証できる[5]．母音などの声

道模型の音響特性は，口唇側に設置したホーンドライバなどから入力した計測信号を声門に設置したマイクロホンで計測することによって得られる[6]．しかし，同様の方法で鼻副鼻腔の声道模型の音響特性の計測は成功しなかった．これは，外鼻孔や鼻腔が狭いためか，声門に設置したマイクロホンで十分な信号対雑音比が得られないためであった．なお，計測信号の音量を大きくすると，声道模型の壁が振動し，かえって信号対雑音比が悪化した．

そこで本研究では，狭い孔から十分な音量の信号を出力できるエクスポーネンシャルホーンと，壁振動を抑制できる厚い壁を持つ声道模型を導入することで，鼻副鼻腔の声道模型の音響特性の計測を試みた．そして，声道模型と同じ鼻副鼻腔の形状データから計算した音響特性と比較検討したので報告する．

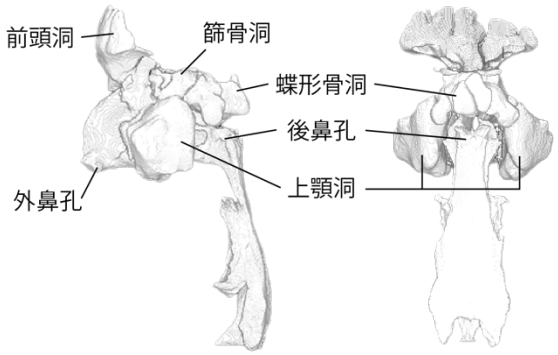


図 1　鼻副鼻腔の形状

2. 材料・方法

2.1. 被験者と CT データ

被験者は両側 ESS III 型鼻中隔湾曲矯正術（前頭洞・前篩骨洞・上顎洞の開放，病的粘膜の処置，鼻中隔湾曲の矯正）を受けた成人男性 1 名である．鼻音/m/発声時の声道形状を CT（SIEMENS SOMATOM Definition Flash）を用いて，空間解像度 0.507×0.507×0.5 mm で撮像した．また，CT 撮像に先立って，別室で仰臥位における/m/の発声を外鼻孔から 2 cm の位置で録音した．なお，本研究は東京慈恵会医科大学附属病院の倫理委員会の承認を受けた（受付番号：30-471（9492））．

2.2. 声道模型の作成

CT 画像を閾値により二値化して体組織と空気を分離し，領域拡大法を用いて声門から外鼻孔に至る鼻副鼻腔を含む声道形状を抽出した．この抽出した声道形状の周囲に壁を付与することで声道模型データを作成した．先行研究では壁の厚さを 3 mm としていたが，本研究では壁振動を抑制するために 5 mm とした．この声道模型データは三角形ポリゴンでデータ形状を表す型式（STL）である．これをナイロン 12 粉末で積層造形して声道模型を作成した．

2.3. 計測方法

図 2 は声道模型の音響特性の計測系の概略を示す．まず，PC で作成したホワイトノイズをスピーカアンプ（BOSE TA-55）で増幅し，エクスポーネンシャルホーン（図 3）[7]で出力する．このエクスポーネンシャルホーンは，全長 114.6 cm で，上端の直径 7 mm の孔からホワイトノイズが出力される．これにより，狭い開口部から高い音圧レベルのホワイトノイズを放射できる．これをマイクロホン（ECM WM-61A 相当品）とポータブル録音機（SONY リニア PCM レコーダーPCM-D10）でサンプリング周波数 48 kHz，量子化ビット数 24 bit，時間長 10 s で録音した（図 2 左）．

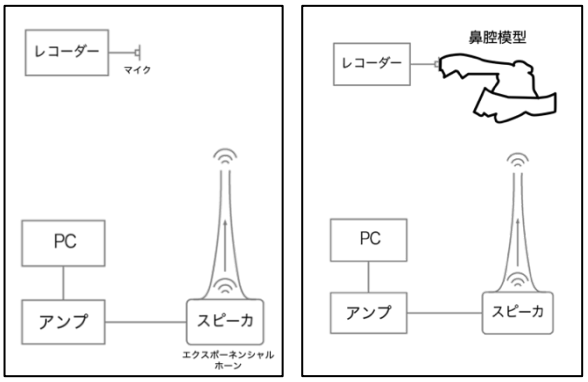


図 2　計測系．左：模型なし，　右：模型あり

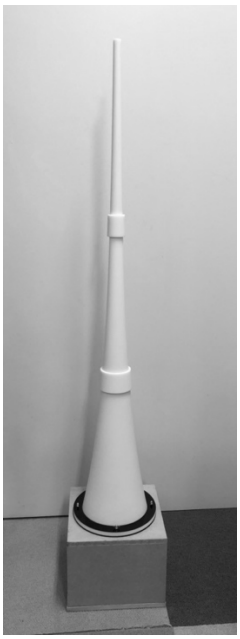


図 3　エクスポーネンシャルホーン

次に，得られた信号から計測系の周波数応答を平坦化するための逆フィルタを作成してホワイトノイズに畳み込んだ．図 4 は，補正前後のホワイトノイズのスペクトルである．補正前は 30 dB 程度あったレベル差が補正後は 3 dB 程度に減少し，ほぼフラットな特性となった．

この補正後の音声を入力信号とし，マイクロホンに声道模型を装着して計測することで，音響の相反定理により声門から外鼻孔までの伝達関数（計測した伝達関数）を得ることができる（図 2 右）.

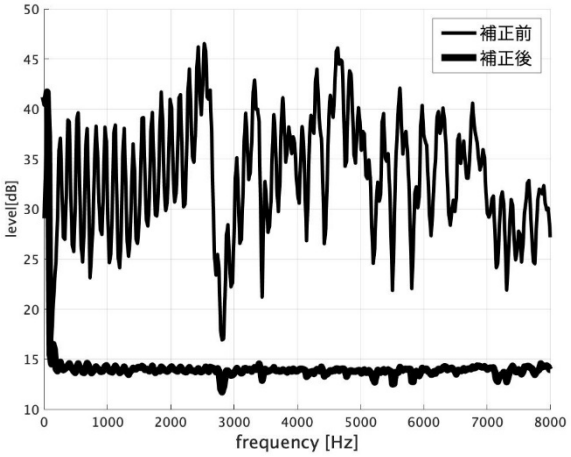


図 4　録音したホワイトノイズ（補正前後）

2.4. 音響シミュレーションによる計算

STL 形式の声道模型データを空間解像度 0.5 mm で離散化し，等方向ボクセルデータに変換して先行研究[4]に従って声門から外鼻孔までの伝達関数（計算した伝達関数）を得た．閉鎖した声門直上に置いた音源点から外鼻孔より下前方 2 cm に置いた観測点における 20 ms のガウシアンパルス応答を FDTD 法で計算した．なお，シミュレーションの周波数は 20 MHz とした．

3. 結果と考察

図 5 は，計算した伝達関数，計測した伝達関数，被験者の音声スペクトルを示す. 計測した伝達関数には，全ての帯域にわたって微細な凹凸がみられたが，声道の共鳴や反共鳴に由来すると思われるピークやディップが存在することから，伝達関数の計測に成功したといえる．なお，この微細な凹凸は，2.3 節で示したホワイトノイズの補正で補正しきれなかった成分であると考えられる．

計算した伝達関数と計測した伝達関数は 8 kHz までの概形が一致した．4 kHz までの 8 つのピーク（P1～P8）のうち，P3，P4 を除く 6 つのピークは，6%以内の誤差で一致した（表 1）．しかし，計算した伝達関数の 2 つのディップ（D1，D2）と，その極零対と考えられる 2 つのピーク（P3，P4）は計測した伝達関数より低域に大きくシフトしていた．

その原因を追究するために，D1，D2 の周波数で声道を励振した際の瞬時音圧分布を計算した．図 6 は D1 周波数における瞬時音圧の分布で，黒いほど音圧の絶対値が大きいことを示す．このディップは，咽頭腔と左前頭洞・右蝶形骨洞が逆相になって振動することによって生成されていた．これは，外鼻孔から放射される音響エネルギーが左前頭洞と右蝶形骨洞で消費されることを示す．すなわち，このディップは左前頭洞と右蝶形骨洞に由来する．図 7 は D2 周波数における瞬時音圧の分布である．図 6 ほど明確ではないが，この周波数では咽頭腔と左上顎洞・右前頭洞が逆相となっていた．すなわち，このディップは左上顎洞と右前頭洞に由来する．

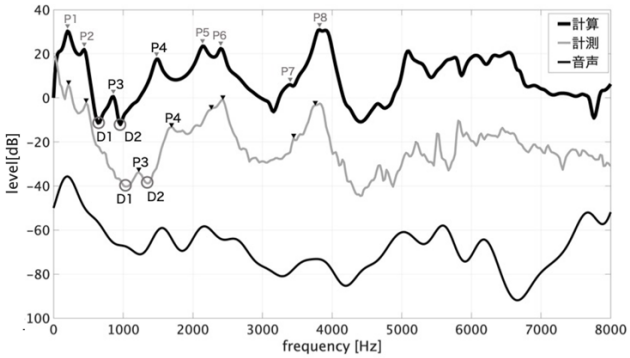


図 5　計算・計測した伝達関数と音声スペクトル

表 1　ピークとディップの周波数の比較

	計算(Hz)	計測(Hz)	誤差(%)
P1	200	210	4.76
P2	440	468	5.98
P3	**860**	**1218**	**29.39**
P4	**1480**	**1687**	**12.27**
P5	2140	2250	4.89
P6	2400	2437	1.52
P7	3448	3440	0.23
P8	3750	3800	1.32
D1	**640**	**1054**	**39.28**
D2	**960**	**1359**	**29.36**

上記の分析から，D1，D2 は副鼻腔に由来するので，副鼻腔の形状によって周波数が変動すると予測される．そこで，声道模型データの体組織と空気を分離する閾値を変化させて伝達関数を計算したところ，D1，D2 が他のピークやディップに比べて周波数やレベルが大きく変化した．これは，閾値操作によって声道全体の形状は変化するが，副鼻腔はヘルムホルツ共鳴器に類似した形状であるため，頭部に相当する鼻腔と連絡する狭い孔はわずかな閾値の変化で相対的に大きな音響変化をもたらすためと考えられる．

これらは，副鼻腔の微細で複雑な形状が精密に造形されていない可能性があることを示唆する．粉末積層

造形は光硬化性樹脂による造形より精度が劣る.また，表面に微細な凹凸が生じやすく，凹部に粉末が残留する場合がある．すなわち，造形におけるこれらの問題が副鼻腔に由来するディップに特に大きな影響を与えていると考えられる.

なお，音声スペクトルは，計算した伝達関数とも計測した伝達関数とも概形があまり一致しなかった．これは，鼻周期や姿勢の変化で短時間の間にスペクトルが変化するにもかかわらず[8]，音声の録音と CT 撮像を同時に実施できなかったためと思われる.

図 6　声道内の瞬時音圧分布の背面観 D1

図 7　声道内の瞬時音圧分布の背面観 D2

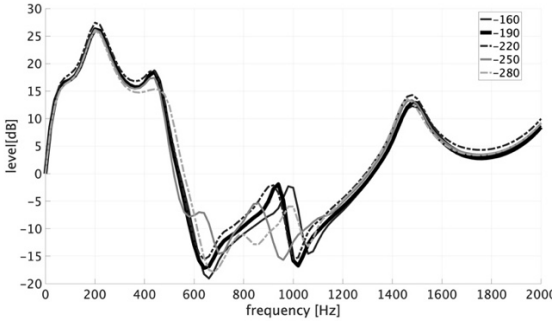


図 8　閾値によるディップの変動

4. 終わりに

本研究では，エクスポーネンシャルホーンを用いて高い音圧レベルの信号を入力し，声道模型の壁を厚くして壁振動を抑制することにより，鼻副鼻腔を含む声道模型の音響特性の計測に成功した．そして，声道模型の形状データから計算した音響特性と比較検討することが可能となった.

比較の結果，計算した伝達関数と計測した伝達関数は 8 kHz までの概形が一致した．また，4 kHz 以下の 8 つのピークのうち 6 つが 6%以内の誤差で一致した．しかし，計算した伝達関数の 2 つの大きなディップと対となるピークは計測した伝達関数の対応するものより周波数が 10%以上低くなった．瞬時音圧分布の分析から，これら 2 つの大きなディップは副鼻腔に由来すること，声道模型データの体組織と空気を分離する閾値を変化させて伝達関数を計算した結果から，副鼻腔の微細構造が正確に造形されていない可能性があることが明らかになった．これを改善するためには，造形法を変更するか，造形する粉末を樹脂に変更するなどを検討する必要がある.

謝　辞

本研究は JSPS 科研費 19K12031 の支援を受けた.

文　献

[1] J. Dang, K. Honda, and H. Suzuki “Morphological and acoustical analysis of the nasal and the paranasal cavities,” J. Acoust. Soc. Am., vol.96, no.4, pp.2088-2100, Oct.1994.

[2] 杉浦唯，竹本浩典，北村達也，内尾紀彦，鴻信義，“鼻音生成時の声道形状の抽出と音響特性の解析，”日本音響学会講演論文集 2020 年 3 月，pp.767-768, Mar.2020.

[3] 杉浦唯，竹本浩典，北村達也，鴻信義，“内視鏡下鼻副鼻腔手術による術前・術後の形状と音響特性の変化の検討，”日本音響学会講演論文集 2021 年 9 月，pp.743-744, Mar.2021.

[4] 杉浦唯，竹本浩典，北村達也，内尾紀彦，鴻信義，“鼻副鼻腔の模擬手術が伝達関数に及ぼす影響，”日本音響学会講演論文集 2021 年 9 月，pp799-800, Spe.2021.

[5] T. Kitamura, H. Takemoto, S. Adachi, and K. Honda, “Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production,” Acoust. Sci. & Tech, vol.30, no.4, pp.288-296, 2009.

[6] H. Takemoto, P. Mokhtari, and T. Kitamura, “Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method,” J. Acoust. Soc. Am., vol.128, no.6, pp.3724-3738, Dec.2010.

[7] J. Epps, J. R. Smith, and J. Wolfe, “A novel instrument to measure acoustic resonances of the vocal tract during phonation,” Meas. Sci. Technol, vol.8, pp.1112-1121, Oct.1997.

[8] 伯田亜海，加地優太，竹本浩典,“鼻周期や首の角度が鼻音の音響特性に与える影響，”日本音響学会講演論文集 2021 年 9 月，pp.795-796, Sep.2021.

音声コミュニケーション環境の対話的試験ツールについて

河原　英紀[†]　榊原　健一[††]　程島　奈緒[†††]　坂野　秀樹[††††]　北村　達也[†††††]
天野　成昭[††††††]

† 和歌山大学 〒 640-8510 和歌山市栄谷 930
†† 北海道医療大学 〒 061-0293 北海道石狩郡当別町金沢 1757
††† 東海大学 〒 108-8619 東京都港区高輪 2-3-23
†††† 名城大学 〒 468-8502 名古屋市天白区塩釜口一丁目 501 番地
††††† 甲南大学 〒 658-8501 神戸市東灘区岡本 8-9-13
†††††† 愛知淑徳大学 〒 480-1197 愛知県長久手市片平二丁目 9
E-mail: †kawahara@wakayama-u.ac.jp, ††kis@hoku-iryo-u.ac.jp, †††hodoshima@tokai-u.jp,
††††banno@meijo-u.ac.jp, †††††t-kitamu@konan-u.ac.jp, ††††††psy@asu.aasa.ac.jp

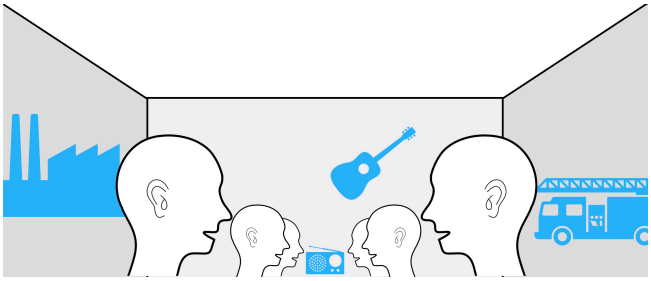
あらまし　音声コミュニケーションは様々な音環境で行われる。音環境は知覚だけではなく発声にも影響を与える。音環境の影響を実験的に調べるためにはそれらを制御可能な手段でシミュレートする必要がある。音環境を対話的に自由に操作する実時間処理を含むツールを実現することで、音声コミュニケーションに関わる音環境の影響についての暗黙知の獲得・蓄積が期待できるだけでなく、それらの影響を定量的に調べることが可能になる。ここでは、予備的な検討を進めているツールの概要と実装について紹介し、議論したい。
キーワード　音声コミュニケーション、実時間処理、対話的環境、音響環境、音声資料収録、音声刺激提示、暗黙知

Interactive test environment for acoustic conditions of speech communication

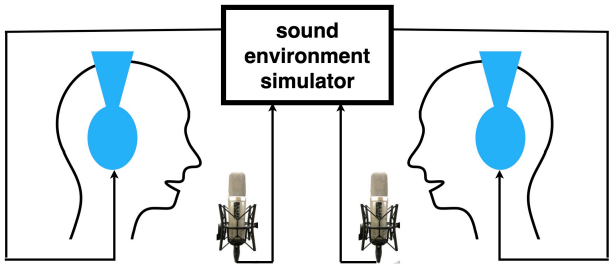
Hideki　KAWAHARA[†], Ken-Ichi　SAKAKIBARA[††], Nao　HODOSHIMA[†††],
Hideki　BANNO[††††], Tatsuya　KITAMURA[†††††], and Shigeaki　AMANO[††††††]

† Wakayama University, 930 Sakaedani Wakayama, Wakayama, 640-8510 Japan
†† Health Sciences University of Hokkaido, 1757 Kanazawa, Tobetsu, Ishikari, Hokkaido, 061-0293 Japan
††† Tokai University, 2-3-23 Takanawa Minato-ku Tokyo, 108-8619 Japan
†††† Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan
††††† Konan University, 8-9-1 Okamoto, Higashinada-ku, Kobe 658-85012, Japan
†††††† Aichi Shukutoku University, Address: 2-9, Katahira, Nagakute-city, Aichi Prefecture, 480-1197 Japan
E-mail: †kawahara@wakayama-u.ac.jp, ††kis@hoku-iryo-u.ac.jp, †††hodoshima@tokai-u.jp,
††††banno@meijo-u.ac.jp, †††††t-kitamu@konan-u.ac.jp, ††††††psy@asu.aasa.ac.jp

Abstract　Physical instantiation of speech communication depends on the acoustic environment. The acoustic environment not only modifies listening behavior but also modifies speech production behavior and makes speech attributes different. Establishing a precisely controllable acoustic environment simulator is necessary to investigate such effects experimentally. Enabling flexible, interactive manipulation of such simulation environment consisting of real-time signal processing will help researchers to efficiently acquire tacit and deep knowledge of speech communication and investigate and quantify their effects by experiments. This paper introduces the preliminary investigations and implementation of such tools to stimulate discussions.
Key words　Speech communication, real-time processing, interactive environment, acoustic environment, speech material acquisition, speech stimuli presentation, tacit knowledge



(a)



(b)

図 1　Speech communication in (a) everyday life condition, and (b) simulated condition using acoustic devices.

1. はじめに

音声によるコミュニケーションは様々な状況で行われる図 1(a)。静かな室内だけではなく、教室や会議など多くの参加者がいる状況や、外の人混みの中などでは、相手の声だけではなく様々な妨害音が同時に聴こえる。また、反響の大きい廊下や拡声器が使われる広い会場では、様々に変形された自分の声も聴こえてくる。合唱などでは、自分の声よりも周りの歌声の方が大きく聴こえることも多い [1]。それらの状況（音環境）に応じて、声の出し方も聴き方も（無意識のうちに、あるいは努力して）変化する [2–6]. これらの影響を客観的に把握し理解することは、様々な状況における音声コミュニケーションの困難を解消するためには必須である [7]。

音声コミュニケーションにおける音環境の影響を実験的に研究しようとする場合には、図 1(b) に示すように音環境のシミュレータを用意して、マイクロフォンで収録した音をヘッドフォンなどで提示する。計算パワーの驚異的な向上（半世紀で 10 億倍 [8]）で信号処理に限ればシミュレータの実現は容易になっている。またシミュレータの構築に必要な情報も手段も急速に整備されてきている [9–13]。

しかし、音環境全体のシミュレーションの実現には、音声の適切な収録と聴覚刺激の適切な提示という物理媒体を介した人間との境界面に多くの問題がある [14–16]。この状況に関しては、『空気伝導型』のイヤフォンの普及(注1)により（適切な状況を設定すれば）問題の多くを回避できる可能性が見えてきた [16]。ここでは、その可能性について説明するとともに、いくつかの予備的な検討について紹介する。

(注1)：通販サイトなどで『空気伝導』『イヤフォン』を検索すると多数がリストされるようになった。(検索時期：2023 年 2 月)

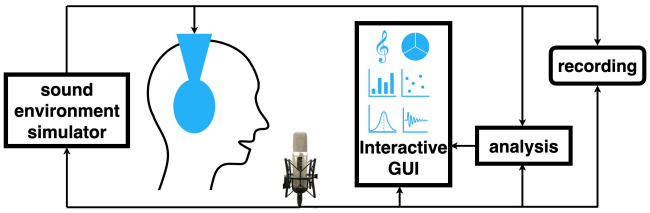


図 2　Feedback with sound environment simulation and interactive visualization.

2. 背　　景

発声の基本周波数に対する聴覚刺激の影響を調べるために、図 2のような実験系 [17] を構成して検討を進めてきた [18]。この実験系で密閉型のヘッドフォンを用いて聴覚刺激を提示した場合、自分の声の聞こえ方（側音）がヘッドフォンをしない場合と大きく異なることが問題となった [16]。この問題を回避するためにヘッドフォンではなくスピーカから聴覚刺激を提示すると、収録された音声に聴覚刺激が混在してしまう。実際、合唱の演奏中の自分の歌声と周囲の歌声のレベルを比較する研究では、気導音だけではなく骨導音の影響を考慮した複雑な処理が必要であった [1]。しかし、それらの処理の誤差や副作用のため、発声される音声の属性を研究する際の収録にこのような手段を用いることは適切ではない。

最近、外耳道を塞がずしかも音漏れを抑圧する仕組みを有するイヤフォン（例えば [19,20] は前述の『空気伝導型』のイヤフォンである）が報告され、入手可能になっている。このようなイヤフォンを用いると、自然な側音に影響を与えずに聴覚刺激を追加するとともに、聴覚刺激を混入させずに音声を収録できる可能性がある [16]。

前述のハードウェアの進歩に加えて、ソフトウェアの進歩も大きい。例えば、科学技術計算用の環境である MATLAB には、対話的な実時間処理を可能にする開発環境（AudioToolbox と Appdesigner）が用意されている [21]。これを用いると、筆頭著者が 1986 年に 2500 行の Pascal で実装した対話的実時間音声分析編集環境 [22, 23] と同等のアプリケーションが、500 行程度の MATLAB コードを書くことで実装できてしまう。この環境では、アプリケーションの動作中にコードを書き換えた結果が即座に動作に反映されるため、対話的アプリケーションの開発の際に特に効果的である。しかも、MATLAB を用いて開発された実時間処理を VST プラグインとして公開することもできる。この環境を用いることで、対話的に高いレベルでの記述により、実行時の経過時間（latency）が少ない実時間処理ができる VST プラグインを開発することができる [24]。

3. 実　装　例

以下では図 2のような実験系の様々な変種の実装例と処理に要する時間の例を示す。用いた環境は、MacBook Pro (macOS Ventura 13.1, M1 Max, 64 GB, MATLAB R2022b Update-4) と Iijima (Windows11 Pro 10.0, Intel Core i7, 4.2 GHz,

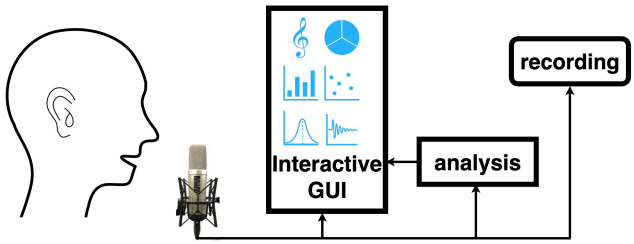


図 3　Feedback with interactive visualization.

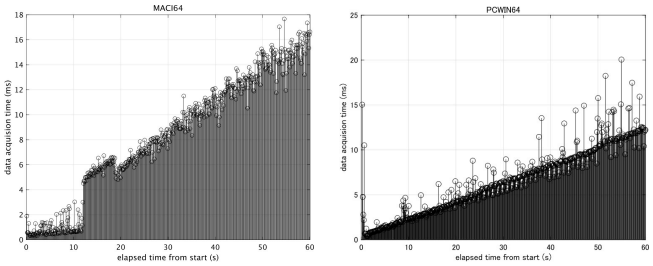


図 4　Elapsed time for transferring captured data using macOS Ventura (left) and Windows11 (right).

16 MB, MATLAB R2022b Update-3) である。開発は macOS で行っており、Windows11 は動作確認用である。

3.1　属性の可視化とフィードバック

まず、フィードバックされる情報に収録された音声を処理した結果を含まない可視化のみの場合を最初に取り上げる (図 3)。ここでは、実時間での音声信号の処理が必要ではないため、AudioToolbox がインストールされていない MATLAB 環境でも開発が可能である。ビデオゲームに関しては 60 fps のフレーム更新速度よりも経過時間が 30 ms 以下である方が重要だとの報告 [25] がある。この結果を歌唱の学習に直接適用できるかは明らかではないが、音環境のシミュレーションに必要となる数 ms のオーダーと比較すると許容できる経過時間は大きい。また、現在の MATLAB では 20 fps を画面更新の上限とする設定(注2)もあり、50 ms 以下で描画を含めて処理を完了することを設計目標としておく。

この属性の可視化に基づくフィードバックでは、入力信号を処理した結果を出力に反映させる実時間処理は不要である。MATLAB のオーディオ録音用のオブジェクト `audiorecorder` は、入力されている信号を非同期に読み出すことが可能なので [26]、描画を更新するループを中心としてアプリケーションを構成することにする。

図 4に、録音時間長と読み出しに要する経過時間を示す。`macOS` も `Windows11` も、録音時間長にほぼ比例して経過時間が増加している。ランダムな変動が含まれているが、最大でも 20 ms 以下であり、描画処理と処理時間に充てる時間に余裕がある。以前報告した歌唱訓練支援用のツールでは、20 秒毎に録音バッファを更新して（1 秒以下の中断の後）録音と描画の更新を再開することとした [27]。

図 5に、描画を更新するループを中心として実装した基本周

(注2)：`drawnow` コマンドの `limitrate` オプション [26]。

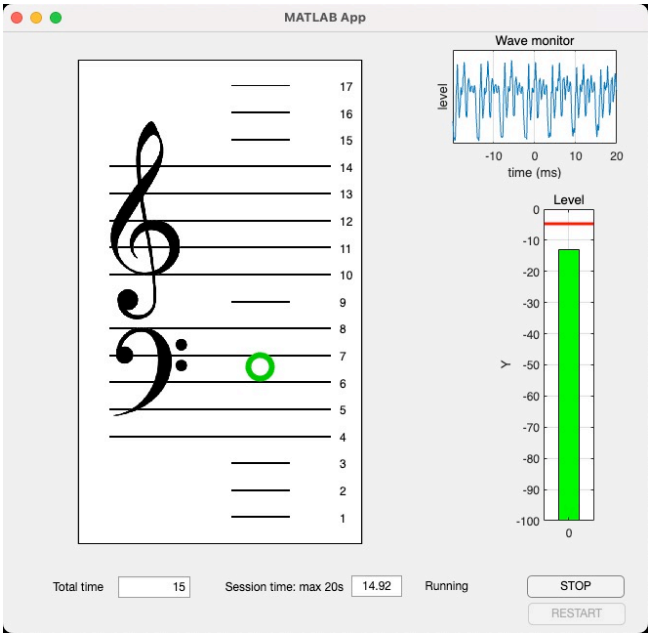


図 5　Snapshot of a visual feedback GUI of voice pitch.

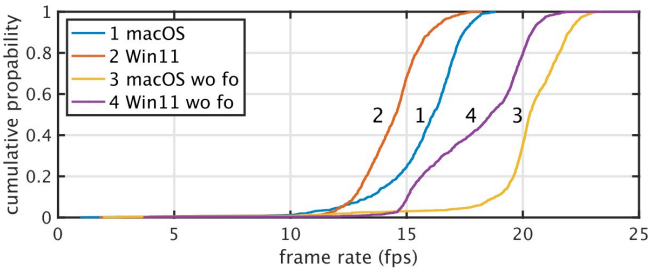


図 6　Display information update rate (fps). “wo fo” represents update without pitch extraction.

波数の可視化 `GUI` の例を示す。描画が終了次第、録音バッファを読み込んで、波形表示、`RMS` 値、瞬時値、基本周波数の表示を更新している。なお、人によっては、声のピッチを『上げて』『下げて』という指示が理解できないことがあるため、五線譜の右に数字を付している [27]。ここでは、以前報告した歌唱訓練支援用のツールのピッチ抽出器を流用した。

図 6に、表示内容・用いた機器の設定の違いによる更新間隔の分布の例を示す。図では、基本周波数の抽出を行わない場合の分布も示している。画面更新の時間間隔がランダムに変動しているため、時々描画が滑らかではなくギクシャクして感じられることがあった。しかし、可視化情報のフィードバックという目的には支障はない。

3.2　対話的可聴化

次に、収録された音声を処理した結果を含まない可聴化を取り上げる（図 7）。ここでも、実時間での音声信号の処理が必須ではない。

図 8にポップアウトボイス（例えば人混みでも目立つ声）研究 [28] の支援のために用意した収録音声確認用のアプリケーションの例を示す。このアプリケーションの目的は、収録された音声のポップアウトの程度を様々な音響環境を対話的切り替

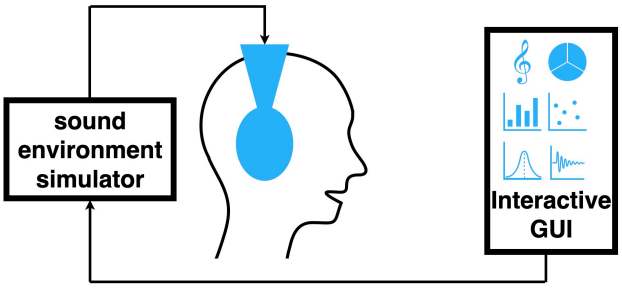
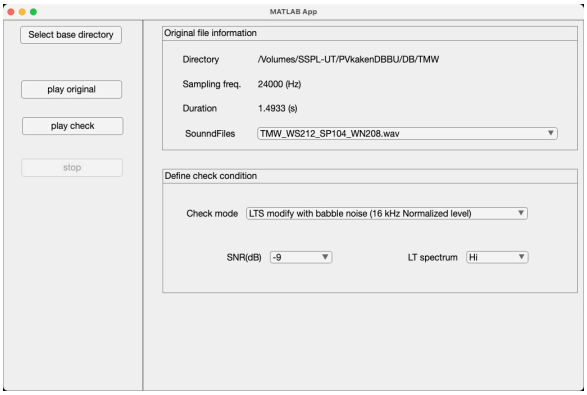
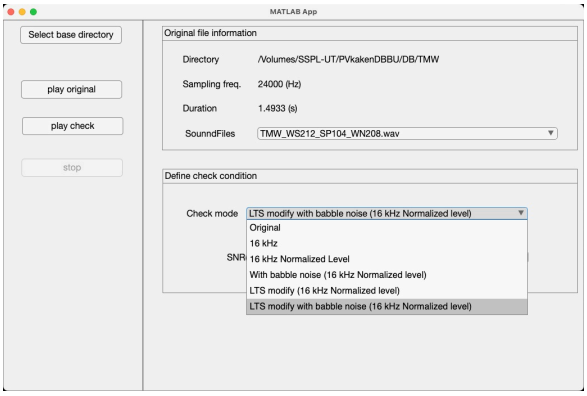


図 7　Evaluation with interactive sonification.



(a)



(b)

図 8　Snapshot of an interactive GUI for pop-out voice study. (a) test mode, (b) condition selection mode.

えながら簡単に確認できる環境の提供である。ポップアウトの程度にはラウドネスが大きく影響するため、Auditory Toolbox のオブジェクトを利用して、ITU の勧告 [29] に基づく正規化を行なっている。音響環境として、現在は（人数をパラメタとした）バブルノイズと SNR およびスペクトル概形（高 PV 話者、中程度 PV 話者、低 PV 話者の概形）を設定できるようにしている。今後は他の種類の雑音や残響など室内音響を追加する予定である。

3.3　音環境のシミュレーション

最後に、収録され処理された音声がフィードバックに含まれる場合を取り上げる（図 2）。この場合、音環境のシミュレーションには短い経過時間で処理結果を出力することが必要になる [30]。

数 ms 程度の短い経過時間が必要な場合には、MATLAB 環境で

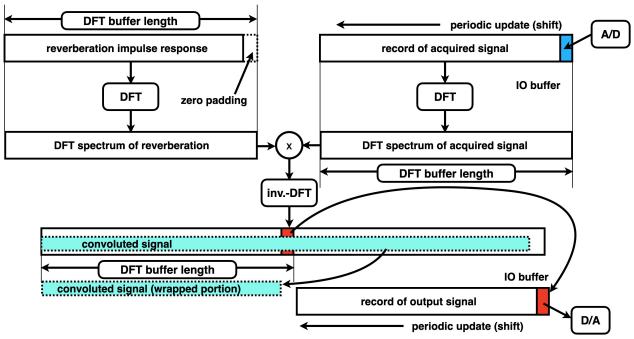


図 9　Low-latency filtering using DFT cyclic convolution.

```
reverveSpectrum = fft(reverbeResponse,fftl);
while currentIteration < nIteration
    inputRecord(1:fftl-frameLength) = inputRecord(frameLength+1:fftl);
    outputRecord(1:fftl-frameLength) = outputRecord(frameLength+1:fftl);
    bufferIn = playrec(bufferOut);
    inputRecord(fftl-frameLength+1:fftl) = bufferIn;
    outputRecord(fftl-frameLength+1:fftl) = bufferOut;
    currentIteration = currentIteration + 1;
    tmp = real(ifft(fft(inputRecord).* reverveSpectrum));
    bufferOut = tmp(fftl-frameLength+1:fftl);
end
```

図 10　Example implementation of this low-latency filtering using DFT cyclic convolution.

はなく VST プラグインなどの実時間処理に適したものを使用しなければならない。しかし、初期反射音が数十 ms 程度の遅れで届く場合には、AudioToolbox で処理できる可能性がある。オーディオインタフェースとの入出力に介在するバッファ長を変更することで、経過時間を短縮することができる。ここでは、MATLAB で室内音響のシミュレーションがどの程度の経過時間で実現できるかを調べる。

室内音響のシミュレーションでは、数秒の長さのインパルス応答を畳み込む必要がある [30]。高速フーリエ変換により演算量を大きく削減できるが、直接的な実現ではインパルス応答以上の長さのバッファを用いる必要があり、経過時間が数秒のオーダーになってしまう。しかし、離散フーリエ変換の性質を利用する [31] と、IO デバイスに介在するバッファ長に対応する時間程度の経過時間で処理することができる。

図 9に仕組みを示した。処理の内容は、図 10に例示した MATLAB による実装と併せると理解しやすい。この実装では、残響のインパルス応答の長さが離散フーリエ変換に用いるバッファ長である 2^{17} サンプルよりも短くなるように設定している（実際には IObuffer の長さとインパルス応答の長さの和を超える長さに、フーリエ変換のバッファ長を設定している [32]。）。離散フーリエ変換（実装は高速フーリエ変換）を用いた畳込みはバッファにある信号が周期的に繰り返される巡回畳み込みになる。収録された信号で埋められた inputBuffer の信号と残響のインパルス応答の（直線）畳み込みの長さは、両者の長さの和になる。この長さは離散フーリエ変換に用いられるバッファ長よりも長いため、周期で折り返されてバッファ上に最初から順に加算される。インパルス応答の長さはこのバッファ長よりも短いため、バッファの最後の部分には折り返しの影響は及ばない。出力に用いられるのは、この影響を受けていない最後の

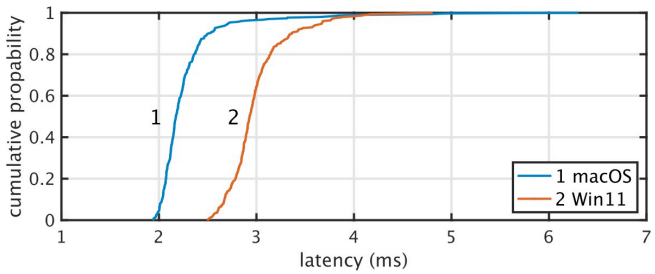


図 11　Measured latencies of filtering using DFT cyclic convolution.

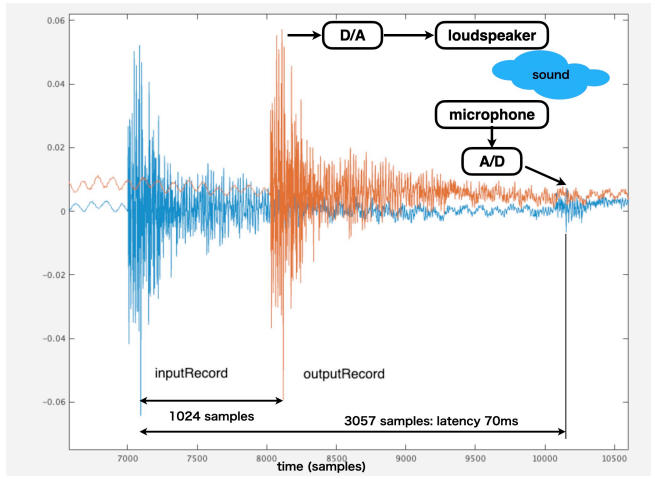


図 12　Measured latencies of filtering using DFT cyclic convolution.

部分（例えば同時入出力用の audioPlayerRecorder 関数の既定値では 1024 サンプル）だけである。結局、この実装では出力に用いられる（例えば 1024 サンプル）以外の計算結果（99% 以上）が毎回廃棄されている。最近のマシンでは CPU 自体に離散フーリエ変換を高速に実行することができる命令が用意されているため、このような無駄が多い実装も許容できる。計算機による離散フーリエ変換の高速化については、文献 [32] の付録に詳しい。

図 11に、1024 サンプルのバッファ長を用いて更新する場合の経過時間を、44100 Hz の標本化周波数として求めた例を示す。macOS と Windows11 の環境で 300 回の測定を行った結果である。他のアプリケーションが動いているマシンの上の MATLAB 環境で測定したため、経過時間は変動している。この経過時間（2 から 3 ms）は、例えば 1024 サンプルに対応する時間長である 23 ms よりも十分に短い。このことは、バッファ長を短くすることができれば、MATLAB 環境でも、特別のハードウェアを使わずに比較的長い残響のある音環境のシミュレーションが実現できることを意味する。さらに短い経過時間が必要な場合には、この処理を VST プラグインとして実装し、優先度の高い実時間処理に向いた構成で利用すれば良い [24]。

3.4　音の入出力を含む経過時間

図 12に、音の入出力を行う系を構成して、音の入力から処理を行った音が記録されるまでの経過時間を測定した例を示す。

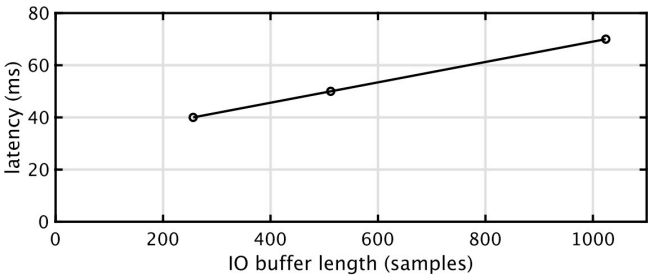


図 13　Relation between IO buffer length and total latency (macOS).

この例では、macOS 上の MATLAB で系を実装した。オーディオインタフェースとして Presonus STUDIO 2|6 USB を用い、音の再生には IK Multimedia の小型アクティブスピーカーである iLound Micro Monitor、マイクロフォンとして AKG C314 ラージダイアフラムコンデンサマイクロフォンを全指向性の設定で用いた。経過時間を読み取りやすいように、小さな金属の文具を叩き合わせた音を用いた。また、実際の実験で使用するインパルス応答の長さ（130047 サンプル）のゼロ詰めした信号の最初に単位インパルスを置いたものを、経過時間測定用のインパルス応答として用いた。

図 12に付記した内容を簡単に説明する。音の入力から処理した信号をインタフェースに送り出すまでの時間は、AudioToolbox の実時間ループの更新時間と一致しており入出力バッファの長さに相当する 1024 サンプルであった。音がマイクロフォンで採取されて処理され、スピーカから音として再生されて再びマイクロフォンで採取されるまでの経過時間は、図中に示すように 3057 サンプルであった。ほぼ 70 ms に相当する。MATLAB を介した処理は、入出力バッファを 2 度経由するため、バッファ長が 1024 サンプルの場合には経過時間は 2048 サンプルに相当する 46.3 ms 以下にすることはできない。

しかし、この例で用いた同時入出力関数（audioPlayerRecorder）の代わりに、入力と出力に別の関数（audioDeviceReader と audioDeviceWriter）を用いることで、バッファ長を設定することができる。入出力のバッファ長を 512 サンプルとした場合には、経過時間は 2197 サンプル（50 ms）となり、256 サンプルの場合には、経過時間は 1770 サンプル（40 ms）となった。ただし、256 サンプルの場合には、入出力を取りこぼすなどのエラーが起こることがあった。図 13に結果を図示してみた。経過時間は入出力のバッファ長に比例する成分と定数成分（OS とハードウェアのドライバなど、MATLAB 以外の要因による成分）から構成されている。MATLAB 環境での処理は、この定数成分と畳み込み処理に要する時間よりも短くすることはできない。（Windows の場合には、ASIO 対応のドライバを使うことで定数成分をより短くできる可能性がある。）

4.　おわりに

ここでは、対話的に音環境を操作するツールについて検討状況と幾つかの具体例を説明した。音声コミュニケーションは様々な音環境で行われる。音環境は知覚だけではなく発声にも

影響を与える。音声コミュニケーションに関わる音環境の影響についての暗黙知の獲得・蓄積と、それらの影響を定量的に調べることを狙って、音環境を対話的に自由に操作する実時間処理を含むツールの検討を進めている。ここで紹介したツールの実装や構成と、それらを開発する狙いなどについて、コメントと議論をお願いしたい。なお、これらの方法やツールは，筆頭著者の GitHub リポジトリで公開している [33]。

謝　　辞

本研究は科研費 20H00291, 21H00497, 21H01596、21H03468、21K19794 の支援を受けた。アルゴリズムなどについて、日頃から議論して頂いている矢田部浩平博士に感謝します。なお、Copilot 機能を組み込んだ新しくなった Windows の Bing に様々な指示をして文章の一部を書換えてみた。元の文章の説明不足などで曲解された部分もあり、文章がどのように受けとられるかを確認する手段としての可能性があると感じた。

文　　献

[1] S. Ternström, D. Cabrera, and P. Davis, "Self-to-other ratios measured in an opera chorus in performance," The Journal of the Acoustical Society of America, vol.118, no.6, pp.3903-3911, 2005. https://doi.org/10.1121/1.2109212.

[2] T.D. Rossing, J. Sundberg, and S. Ternström, "Acoustic comparison of voice use in solo and choir singing," The Journal of the Acoustical Society of America, vol.79, no.6, pp.1975-1981, 1986. https://doi.org/10.1121/1.393205.

[3] W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," The Journal of the Acoustical Society of America, vol.84, no.3, pp.917-928, 1988. https://doi.org/10.1121/1.396660.

[4] M. Södersten, S. Ternström, and M. Bohman, "Loud speech in realistic environmental noise: Phonetogram data, perceptual voice quality, subjective ratings, and gender differences in healthy speakers," Journal of Voice, vol.19, no.1, pp.29-46, 2005. https://doi.org/10.1016/j.jvoice.2004.05.002.

[5] K. Reid, P. Davis, J. Oates, D. Cabrera, S. Ternström, M. Black, and J. Chapman, "The acoustic characteristics of professional opera singers performing in chorus versus solo mode.," Journal of Voice: Official Journal of the Voice Foundation, vol.21, no.1, pp.35-45, 2006. https://doi.org/10.1016/j.jvoice.2005.08.010.

[6] T. Sierra-Polanco, L.C. Cantor-Cutiva, E.J. Hunter, and P. Bottalico, "Changes of voice production in artificial acoustic environments," Frontiers in Built Environment, vol.7, no.666152, pp.1-13, 2021. https://doi.org/10.3389/fbuil.2021.666152.

[7] 赤木正人, "確実に情報を伝える音声避難誘導システムの構築に向けて," 日本音響学会音声研究会, vol.3, no.1 SP-2023-5, pp.29-28, 2023.

[8] C.E. Leiserson, N.C. Thompson, J.S. Emer, B.C. Kuszmaul, B.W. Lampson, D. Sanchez, and T.B. Schardl, "There's plenty of room at the top: What will drive computer performance after moore's law?," Science, vol.368, no.6495, p.eaam9744, 2020. https://doi.org/10.1126/science.aam9744.

[9] 伊勢史郎, "没入型聴覚ディスプレイ「音響樽」," 日本バーチャルリアリティ学会誌, vol.25, no.2, pp.7-12, 2020. https://doi.org/10.18974/jvrsj.25.2_7.

[10] V. Hohmann, R. Paluch, M. Krueger, M. Meis, and G. Grimm, "The virtual reality lab: Realization and application of virtual sound environments," Ear and Hearing, vol.41, no.Suppl 1, p.31S, 2020. https://doi.org/10.1097/AUD.0000000000000945.

[11] S. Liu and D. Manocha, "Sound synthesis, propagation, and rendering: A survey," 2020. https://arxiv.org/abs/2011.05538.

[12] M. Vorländer, "Are virtual sounds real?," Acoustics Today, vol.16, no.1, pp.46-54, 2020. https://doi.org/10.1121/AT.2020.16.1.46.

[13] T.W. Leishman, S.D. Bellows, C.M. Pincock, and J.K. Whiting, "High-resolution spherical directivity of live speech from a multiple-capture transfer function method," The Journal of the Acoustical Society of America, vol.149, no.3, pp.1507-1523, 2021. https://doi.org/10.1121/10.0003363.

[14] R.R. Patel, S.N. Awan, J. Barkmeier-Kraemer, M. Courey, D. Deliyski, T. Eadie, D. Paul, J.G. Švec, and R. Hillman, "Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function," American J. Speech-Language Pathology, vol.27, no.3, pp.887-905, 2018. https://doi.org/10.1044/2018_AJSLP-17-0009.

[15] 榊原健一, 河原英紀, 水町光徳, "利用価値の高い音声データの録音手順," 日本音響学会誌, vol.76, no.6, pp.343-350, 2020. https://doi.org/10.20697/jasj.76.6_343

[16] H. Kawahara, K.-I. Sakakibara, M. Mizumachi, M. Morise, and H. Banno, "Comparative measurement of headphones using new test signals and side tones while voicing," Proc. Auditory Res. Meeting, vol.52, no.8 H-2022-104, pp.575-580, 2022.

[17] H. Kawahara, T. Matsui, K. Yatabe, K.I. Sakakibara, M. Tsuzaki, M. Morise, and T. Irino, "Implementation of interactive tools for investigating fundamental frequency response of voiced sounds to auditory stimulation," Proc. APSIPA ASC, pp.897-903, 2021. http://www.apsipa.org/proceedings/2021/pdfs/0000897.pdf.

[18] 廖嘉慧, 河原英紀, 松井淑恵, "基本周波数の周波数変調に対する発声の不随意応答: 刺激音の種類・変調量による影響," 日本音響学会秋季研究発表会, 3-P-23, pp.967-972, 2022.

[19] 加古達也, 千葉大将, 小林和則, "オープンイヤー型イヤホン向け開口エンクロージャ構造のシミュレーション評価の検討," 日本音響学会秋季研究発表会, 1-R-11, pp.415-416, 2022.

[20] 千葉大将, 加古達也, 小林和則, "音漏れ低減のためのオープンイヤー型イヤホン向け 開口エンクロージャ構造の提案," 日本音響学会秋季研究発表会, 1-R-12, pp.417-418, 2022.

[21] The Mathworks, Inc., Natick, MA USA, "MATLAB R2022b," 2022.

[22] 河原英紀, "音声知覚過程研究支援環境のユーザインタフェース," 日本音響学会聴覚研究会資料, vol.H-87, no.21, pp.1-7, 1987.

[23] NTT アドバンステクノロジ, "音声工房," 1989. (発売した年を表記。Turbo Pascal を用いて実装されたのは 1986 年).

[24] 岩村　宏, MATLAB で簡単オーディオ プラグイン開発: エフェクター・プラグインが数分〜で作れる! (VST/AU), Amazon, 2023. (著者による note には MATLAB を用いた信号処理に関する多数の有用な資料が掲載されている。) https://note.com/leftbank/m/m088894fecb57.

[25] J. Spjut, B. Boudaoud, K. Binaee, J. Kim, A. Majercik, M. McGuire, D. Luebke, and J. Kim, "Latency of 30 ms benefits first person targeting tasks more than refresh rate above 60 Hz," SIGGRAPH Asia 2019 Technical Briefs, pp.110-113, 2019.

[26] The Mathworks, Inc., Natick, Massachusetts, "ヘルプセンター," 2023. (MATLAB の技術情報を検索できる) https://jp.mathworks.com/help/index.html.

[27] H. Kawahara, K.-I. Sakakibara, E. Haneishi, and K. Hagiwara, "Real-time and interactive tools for vocal training based on an analytic signal with a cosine series envelope," Proc. APSIPA ASC, pp.907-910, 2019. https://doi.org/10.1109/APSIPAASC47483.2019.9023094.

[28] T. Kitamura, N. Kunimoto, H. Kawahara, and S. Amano, "Perceptual Evaluation of Penetrating Voices through a Semantic Differential Method," Proc. Interspeech 2022, pp.3063-3067, 2022. https://doi.org/10.21437/Interspeech.2022-100.

[29] EBU R 128, "Loudness Normalisation and Permitted Maximum Level of Audio Signals," European Broadcasting Union, 2014. https://tech.ebu.ch/docs/r/r128.pdf.

[30] 伊勢史郎, "音場共有を実現するための低遅延・実時間信号処理とその実装," 日本音響学会誌, vol.73, no.9, pp.608-614, 2017. https://doi.org/10.20697/jasj.73.9_608.

[31] 河原英紀, "ディジタル信号処理の落とし穴," 日本音響学会誌, vol.73, no.9, pp.592-599, 2017. https://doi.org/10.20697/jasj.73.9_592.

[32] 矢田部浩平, "第二回：離散フーリエ変換," 日本音響学会誌, vol.77, no.5, pp.331-338, 2021. https://doi.org/10.20697/jasj.77.5_331.

[33] H. Kawahara, "GitHub repository for speech and hearing research/education tools," 2023. (retrieved 31 Jan. 2023). https://github.com/HidekiKawahara

2023年3月3日, 日本音響学会音声コミュニケーション研究会資料, Vol. 3, No.2, SC-2023-9
2023年3月3日, 日本音響学会音声研究会資料, Vol. 3, No.2, SP-2023-12

一語発話「ん」を用いた日本語の感情表現の韻律特徴
－日本語母語話者による予備調査の結果－

レレトゥン† 島村徹也† 鮮于 媚‡

†埼玉大学、大学院理工学研究科 〒338-8570 埼玉県さいたま市桜区下大久保 255
‡埼玉大学、大学院人文社会科学研究科 〒338-8570 埼玉県さいたま市桜区下大久保 255
E-mail: †lae.lae.h.378@ms.saitama-u.ac.jp, ‡sonumee@mail.saitama-u.ac.jp

あらまし: 本研究は日本語の感情表現時の韻律特徴を明らかにし、第ニ言語習得時に応用することを目的とし、一語発話「ん」を用いた感情表現を調査した。調査は異なる文脈情に基づき発話をし、収録した。収録した音声データの F0 の値を分析した。結果は、F0 のダイナミックレンジが「Rise and Fall」である場合は、肯定的な感情表現で、「Gradual Fall」は、否定的な感情表現、疑いの表現では、「Rise」である可能性が示唆された。今後、日本語学習者を含む知覚および生成調査を行い、検証する予定である。
キーワード：感情表現、韻律特徴、一語発話“ん”、異文化コミュニケーション、パラ言語

Prosodic Features of Japanese Emotional Expressions using One-word Utterance “n”
－Results of a preliminary survey by native Japanese speakers－

Lae Lae Htun,† Tetsuya SHIMAMURA†, and Mee SONU‡

†Graduate School of Science and Engineering, Saitama University 255 Shimookubo, Sakura-ku, Saitama-shi, Saitama, 338-8570 Japan
‡Graduate School of Humanities and Social Sciences, Saitama University 255 Shimookubo, Sakura-ku, Saitama-shi, Saitama, 338-8570 Japan

E-mail: †lae.lae.h.378@ms.saitama-u.ac.jp, ‡sonumee@mail.saitama-u.ac.jp

Abstract: Our aim was to understand the scientific features of emotional expressions in Japanese by native and non-native speakers. Specifically, this study focused on how non-native speakers recognize the emotional expression of native speakers and vice versa. As a first attempt, the present study analyzed various emotional expressions using the one-word utterance “n” by young female native Japanese speakers. Based on the preliminary survey, the results of the analysis showed three types of F0 dynamic patterns in “n.” Positive and agreeable emotions of “n” exhibited the “Rise and Fall” pattern, negative emotions of “n” exhibited the “Gradual Fall” pattern, whereas doubtful emotions of “n” exhibited the “Rise” pattern. The minimum frequency of F0 did not considerably differ in all these emotions, whereas the maximum frequency of F0 was high for all emotions except negative emotions. These results suggest that the F0 movement may be related to emotional expression, that is, positive and negative.
Keywords: Emotional expression, Prosodic features, One-word utterance "n", Cross-cultural communication, Paralinguistics

1. Introduction

A speech signal that does not have linguistic content itself but can transfer plenty of information with phonetic content is usually classified as “paralinguistic” information [1]. Campbell (2004) investigated paralinguistic information using 129 utterances of the interjection "eh" from conversational speech. The results showed that voice-quality information was an important discriminator and could express a variety of meanings or pragmatic effects in conversation. In the second group,

annotation was examined, and two types of patterns in annotating speech were determined: categorical and dimensional.

In the categorical annotation, various labels, such as anger, sadness, happiness, fear, and disgust, are assigned for emotion annotation. An appropriate label is annotated to utterances or segments of speech, which could be useful in making AI communicative agents more realistic [4]. Kessensa (2009) made both categorical and dimensional annotations for neutral and emotional speech synthesis (anger, fear, sad, happy, and relaxed). The results showed that classification rates were higher for sentences where linguistic meaning matched with emotion.

In dimensional annotation, each utterance or segment is defined with the scopes of the dimensions and expressed as a point in the paralinguistic space. Dimensional annotation is relevant to describe paralinguistic information in spontaneous speech [4]. Greenberg (2009) investigated paralinguistic prosody control in conversational speech impressions using multi-dimensional scaling. The results showed that three-dimensional perceptual impressions were manifested, and each dimension was correlated with different F0 characteristics. The positive-negative impression could be controlled by average F0 height, whereas confident-doubtful or allowable-unacceptable impressions can be controlled by F0 dynamic patterns.

Mokhtari (2003) focused on the relation between paralinguistic information and acoustic features of Japanese monosyllabic words, which are typically backchannel and filler utterances in Japanese. The experiment used 141 utterances of “hai,” “un,” and “ah” from spontaneous conversational speech recordings of one Japanese female native speaker. The result showed that prosodic features affected the perceptual impressions of different utterance types.

Although most studies have focused on the perceptual impressions of paralinguistic information in conversational speech, the present study used conscious speech intentionally generated by native Japanese speakers playing card games. Unlike conversational speech, Japanese card games have no conversational content and players guess each other’s emotions by only listening to their utterances. There are many theme cards in Japanese card games, such as “n,” “eh,” “haa,” “hee,” “uso,” and “maji.”

The present study used one theme card “n” because it is a good sample to observe the prosodic features of one-word utterances without any other linguistic and conversational content [11]. The Japanese word “n” is frequently used in informal speech to express different kinds of emotions and attitudes. It is an interjection, rejoinder, or filler word, depending on the context or situation. Especially, “n” is a sustained utterance, and it does not have any lexical meaning, default intonation, or accent. However, its F0 pattern transfers considerable paralinguistic information [2]. Based on the experimental result of Greenberg et al. (2009), we examined various patterns of prosodic features in one-word utterance “n,” such as different F0 dynamic patterns, F0 average height, and duration corresponding to the speaker’s emotions and intentions.

2. Experiments

2.1 Participant

11 female speakers of native Japanese were consciously generated the speech of “n” with 8 different types of emotion expressions. Mean of the age was 21.6 years for the speakers and they all spoke Tokyo dialect.

Each of the participant were given a pile of “n” theme cards from Japanese card games [11]. The 8 different specified emotion expressions on “n” card were feeling of (A) when entering smooth and comfortable futon (フカフカお布団に入って「ん」), (B) deep thought with unhappiness（考え込んで「ん」）, (C) agreement（納得の「ん」）, (D) tolerance（我慢して「ん」）, (E) disgusting（うんざりして「ん」）, (F) doubtful（ちょっと違うなの「ん」）, (G) unwilling agreement（しぶしぶ同意の「ん」）, (H) delicious（おいしいの「ん」）.

All emotion expressions were recorded at least 3 times and after the recording process, speakers were allowed to listen to their own speech samples and can record again if they needed.

Overall, 88 “n” one-word utterances were recorded in a soundproof room using a TASCAM PCM recorder.

2.2 Speech Analysis Method

For the experiment and analysis, we used speech analysis software Praat and Microsoft Excel. First, we used the Praat script to extract the F0 characteristics, such as pitch-listings, F0-Min, F0-Max, F0-average height, and duration of the one-word utterance “n.” After obtaining F0 characteristics, we divided the pitch-listings into five equal parts and then plotted the F0 contour using Microsoft Excel.

Next, the authors classified each contextual expression, from A to H, according to the meaning of the sentence. A and H were classified as positive emotions, C as agreeable, B, D, G, and E as negative emotions, and F as doubtful.

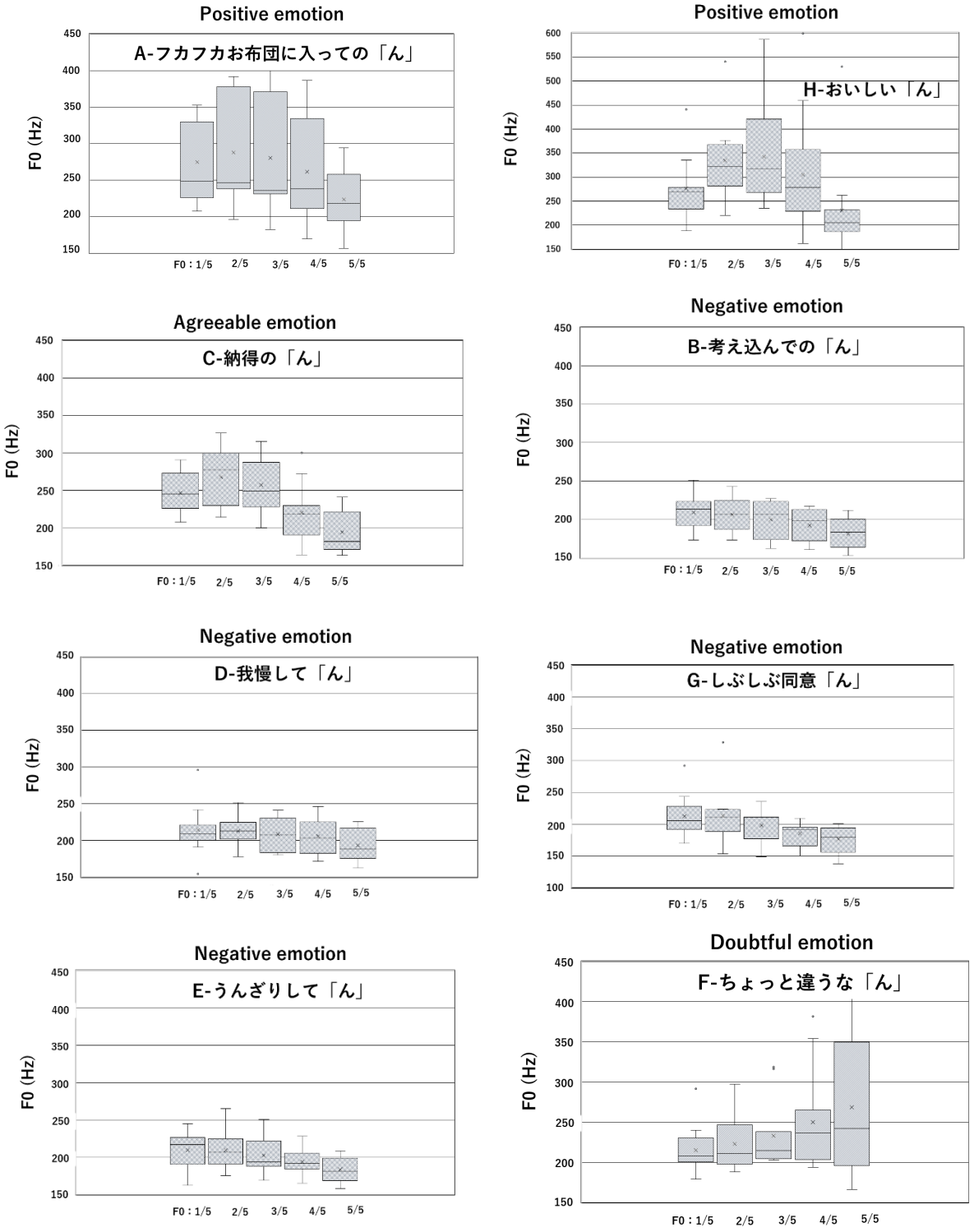


Fig. 1 F0 (Hz) dynamic patterns of one-word utterance "n" with 8 different kinds of emotions.

3. Results

3.1 F0 contour of utterance

Fundamental frequency (F0) contours of 8 different emotional utterances are plotted in Fig.1. F0 contour in positive emotion (A and H) and agreeable (C) emotion "Rise & Fall" patterns. Negative emotion utterances are B, D, E, G and their F0 contours were "Gradual Fall" patterns. And F0 contours in doubtful emotion utterances of F were "Rise" pattern.

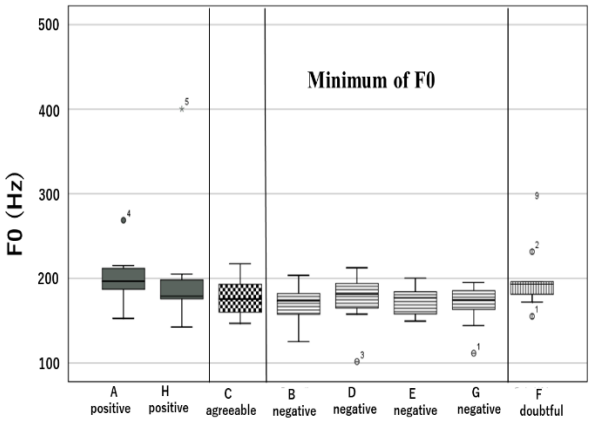


Fig.2 The minimum value of the F0

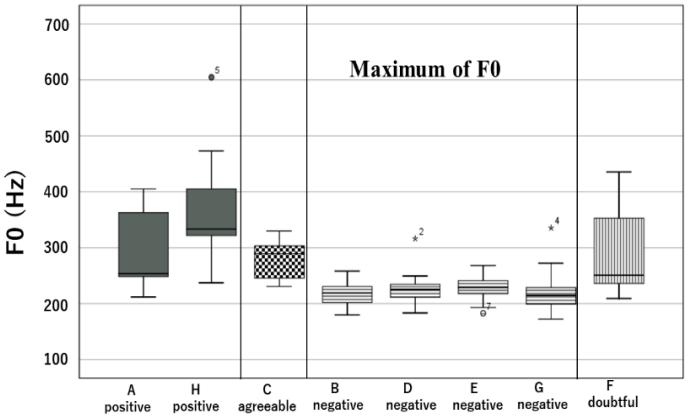


Fig.3 The maximum value of the F0

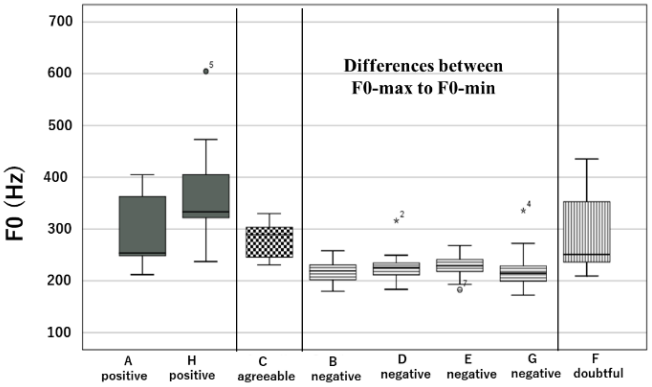


Fig.4 Differences between maximum F0 and minimum F0

3.2 The prosody features of utterances

In each context, we measured the minimum and maximum value of the F0 (Fig.2 and Fig.3). Fig. 2 shows that the value of F0 is low for all utterances. There are no remarkable differences the between emotions conveyed.

Fig. 3 shows that the value of F0 is high for A and H, whereas negative emotions B, D, E, and G have low values of F0. These tendencies also reflect the differences between maximum F0 and minimum F0 (Fig. 4).

These tendencies indicate that the positive emotions (A and H) have a large dynamic range of F0. In contrast, the negative emotions (B, D, E, G) have a small dynamic range of F0. F0 is one of the factors in expressing emotions in Japanese. Moreover, the dynamic range of F0 is also an objective indicator used to understand the emotional features of utterances.

3.3 Duration

We measured the duration of the one-word utterance “n” (Table 1). Table 1 shows the duration of the one-word utterance “n” for each emotion. There seems to be no common characteristic difference between the emotions. Moreover, there are large differences between speakers.

Therefore, it should be investigated further to obtain a more in-depth understanding. The mean of the duration shows that negative emotions B and D have longer durations than other emotions, whereas agreeable positive emotion A has the shortest.

Table 1. Duration of one-word utterance “n” with 8 different kinds of emotions, uttered by 11 Japanese native speakers.

Context / Speaker	8 Different kinds of Emotions (ms)							
	Positive		Agreeable		Negative			Doubtful
	A	H	C	B	D	E	G	F
N001	900	500	550	1350	1050	1000	530	1090
N002	880	820	400	710	520	690	530	680
N003	740	450	360	900	1320	410	470	680
N004	740	370	530	1000	860	370	1070	480
N005	680	730	450	770	1190	640	390	930
N006	580	460	280	640	560	460	380	620
N007	760	520	710	820	550	820	470	750
N008	680	400	260	610	680	600	580	700
N009	760	830	690	820	870	690	730	550
N010	800	620	610	740	590	610	580	680
N011	490	390	380	690	700	480	400	620
Average Duration	728	554	475	823	808	616	557	707

4. Conclusion

In this study, we examined the F0 characteristics of various emotional expressions using one word utterance “n” in Japanese using a part of card games. The results showed that there were 3 types of F0 dynamic patterns; “Rise & Fall” patterns occurred in positive and agreeable emotions of “n”, “Gradual Fall” patterns occurred in negative emotions, and “Rise” patterns occurred in doubtful emotions. Minimum frequency of F0 is not remarkably different in all these emotions while the maximum frequency of F0 is high for all emotions except negative emotions. Average F0 heights were high in positive and agreeable emotions. Duration of agreeable emotion (C) is the shortest, conversely, duration of negative emotions (B, D, E, G) are longer than other emotions. Duration of positive emotions are similar to negative emotions. These F0 characteristics findings will be useful for prosody learning of Japanese. In addition, it is necessary to proceed with the detailed analysis of duration, speaker’s attitudes, and perceptual impression test among interactive senses.

5. Future plan

In this study, we investigated how Japanese native speakers express emotional differences while using the one-word “n” utterance. The results show that the dynamic range of F0 could be the indicator to identify the expression of emotions. These results encourage us to investigate how non-native perceive the one-word “n” utterance and how non-native speakers generate emotional expressions. Next, we are going to investigate not only the one-word “n” but also word expressions.

Acknowledgement

This research was supported by Saitama University’s Shimamura Research Lab and Sonu mee Research Lab. We are deeply grateful to everyone who cooperated and participated in the research survey processes.

References

[1] H. Mori, T. Satake, M. Nakamura, H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics”, Speech Communication 53, pp. 36–50, 2011.

[2] Y. Yamashita, “A review of paralinguistic information processing for natural speech communication”, Acoust. Sci. & Tech. 34, pp. 73-79, 2013.

[3] Y.Greenberg, N.Shibuya, M. Tsuzaki, H. Kato, Y. Sagisaka, “Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling”, Speech Communication 51, pp. 585–593, 2009.

[4] N. Campbell, D. Erickson, “What do people hear? A study of the perception of non-verbal affective information in conversational speech”, J. Phonet. Soc. Jpn. 8 (1), pp. 9–28, 2004.

[5] Judith M. Kessens, Mark A. Neerincx, Melanie Kroes, Gerrit Bloothooft, Rosemarijin Looije, “Perception of synthetic emotion expressions in speech: Categorical and dimensional annotations”, IEEE 3rd Int. Conf. on Affective and Intelligent Computing, Amsterdam, p. 448-452, 2009.

[6] A. Iida, P. Mokhtari and N. Campbell, “Acoustic correlates of monosyallabic utterances of Japanese in different speaking styles”, 15th International Congress of Phonetic Sciences, Barcelona, 2003.

[7] M. Celce-Murcia, Donna M. Brinton, Janet M.

Goodwin, “Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages”. Cambridge University Press, UK, 1996.

[8] BabySparks (Social-Emotional), “What is Prosody & Why is it Important?”, Retrieved 20 August 2022 from https://babysparks.com/2020/05/14/what-is-prosody-why-is-it-important/

[9] P. A. Abhang, B. W. Gawali, & S. C. Mehrotra, “Technical Aspects of Brain Rhythms and Speech Parameters”, Introduction to EEG and Speech-Based Emotion Recognition, pp 51–79, 2016.

[10] M. Kjeldgaard, “Prosodic Features of Japanese and English”, Language and Culture No. 34, Aichi University Press, Japan, 2016.

[11]『はぁって言うゲーム』幻冬舎

[12] Praat http://www.fon.hum.uva.nl/praat/

女性声優の声質表現語抽出の試み

安田 茉　北村 達也

甲南大学知能情報学部　〒 658-8501 兵庫県神戸市東灘区岡本 8–9–1

E-mail: t-kitamu@konan-u.ac.jp

あらまし アニメーション作品やゲームのキャラクタとその声を担当する声優の声質とのミスマッチによる違和感を減らすため，女性声優の声質を表現する語を抽出した．女性声優の声質を表す様々な表現語を収集し，それらの了解性，同義性，類似性を調査し，クラスタ分析によって「大人っぽい声–幼い声」,「上品な声–荒々しい声」,「優しい声–冷たい声」,「元気な声–おしとやかな声」,「高い声–低い声」の 5 対の声質表現語対を得た．

キーワード アニメーション作品，ゲーム，キャラクタ，声優，声質，クラスタ分析

Extraction of expressions associated with voice quality of female cartoon voice actors

Matsuri Yasuda and Tatsuya Kitamura

Faculty of Intelligence and Informatics, Konan University　8–9–1, Okamoto, Higashinada, Kobe, Hyogo, 658-8501, Japan

E-mail: t-kitamu@konan-u.ac.jp

Abstract Expressions associated with the voice quality of female cartoon voice actors were extracted to reduce mismatch between their voice quality and cartoon characters of animation films and video games. We first collected expressions describing the voice quality, and then investigated their understandability, synonymity, and similarity. Five expression pairs were extracted through a cluster analysis.

Keywords Animated cartoon, Video games, Character, Voice actors, Voice qualities, Cluster analysis

1　はじめに

現在，日本では毎年多数のアニメーション作品やゲームが制作され，数多くのキャラクタが生み出されている．声優に関する雑誌 [1] によれば，それらの声を担当するプロの声優は 2022 年 3 月の時点で 1,658 名おり，それぞれが固有の声質を 1 つ以上持っている．

声優のキャスティングはキャラクタの印象を決定づける．原作のキャラクタの性格やイメージとアニメーション作品の声優の声質との間にミスマッチがあれば，視聴者が違和感を感じることがある．アニメーション作品の制作発表後や公開後にファンの間でキャスティングに対する意見が分かれ，議論になることも少なくない．声優のキャスティングは時に作品の商業的な成否をも左右する重要な要因である．

本研究では，キャラクタと声優の声質のミスマッチによる違和感を減らすため，女性声優の演技音声の声質表現語について体系的な調査を行う．本研究にて対象を女性声優に限定した理由は，女性の方が声質の幅が広く，多くの声質表現語を抽出できると考えたためである．

一般の話者による通常発話の声質を表す日常表現語については，木戸と粕谷 [2, 3] が体系的な検討を行っている．彼らの 1999 年の報告 [2] では，ま

表 1: 木戸と粕谷 [3] により抽出された声質表現語.

表現語対		
高い声	–	低い声
かすれた声	–	澄んだ声
落ち着きのある声	–	落ち着きのない声
迫力のある声	–	弱々しい声
太い声	–	細い声
張りのある声	–	張りのない声
表現語 (反意語を持たない)		
鼻声		

ずアンケートと文献調査により声質に関連する 137 語を選定した．次に，表現語の了解性と同義性に関するアンケートを実施し，25 語を抽出した．その 25 語を用いて実験参加者自身の音声を評価する手法 (自己評価法) によって表現語間の類似性を調査し，クラスタ分析により 10 語の表現語を抽出した．それらの表現語から 8 対の表現語対と 1 つの表現語を得た．この研究に続く彼らの研究 [3] では，上記の自己評価法の部分を聴取実験にて実施し，その結果，表 1 に示す 6 対の表現語対と 1 つの表現語を得た．

本研究では，木戸と粕谷 [2, 3] の手法を踏襲し，女性声優の演技音声を対象に声質表現語対を抽出する．得られる声質表現語対を用いて女性声優の声質を統一的に表現することができれば，制作者の意図に合ったキャスティングの一助となることが期待される．

2 声質表現語の収集

2.1 実験協力者

アニメーションや声優が好きな大学生 4 名 (男女各 2 名) が参加した．

2.2 方法

実験協力者に，女性声優の声質を表す表現語を 50 語から 100 語を目安に列挙させた．作業は 4 名独立に行わせた．収集した表現語から重複を削除し，木戸と粕谷 [2] と同様に以下に該当する語を除外した．

1. 声質を表すとは言えない語 (「大きな」など)
2. 声質を表しているが，特定の場面で用いられる語 (「ひそひそした」,「ささやいた」など)
3. 主として発話者の心理状態を表す表現語 (「怖い」,「寂しい」など)
4. 他の表現語の否定語 (「淀みのない」,「品のない」など)
5. 他の表現語の明らかな同義語 (「色っぽい」に対する「セクシーな」など)

ただし，本研究は女性声優の演技音声に関する声質表現語を抽出するため,「はきはきした」などの声質を表しているとは言えない表現語や聴取者の好悪，心理状態を表す表現語に該当していても必要と感じられるものは残した．

2.3 結果

のべ 351 語が収集され，重複した語を削除した結果，282 語の表現語が得られた．さらに，上記 1 から 5 に該当する 26 語を除去した結果，256 語となった．

3 声質表現語の了解性の調査

3.1 実験協力者

大学生 67 名が調査に参加した．

3.2 方法

前節にて得られた表現語 256 語のそれぞれが以下のいずれにあてはまるかを二肢強制選択させた．

(a) 実際に使用したことがあるか，あるいは使わないがどのような音声を指すのか想像できる語

(b) 使ったことも聞いたこともないか，聞いたことはあるがどのような音声を指すのかわからないか，あるいは意味が分からない語

全体の 90%以上の実験協力者が (a) を選択した語を了解性が高い語とした．

3.3 結果

上記の調査の結果，40 語が抽出された．

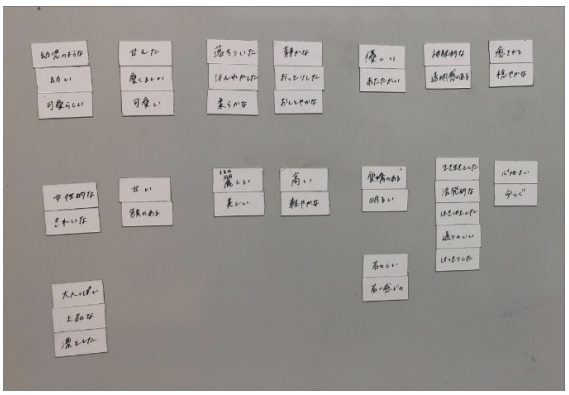


図 1: 同義性の調査の例.

4 声質表現語の同義性の調査

4.1 実験協力者

大学生 8 名 (男性 4 名，女性 4 名) が調査に参加した.

4.2 方法

20 mm×40 mm のマグネットシートを 40 枚用意し，前節にて抽出された 40 語を 1 枚につき 1 語ずつ記入した．実験協力者には，図 1 に示すように同義と考えられる表現語をまとめてホワイトボードに貼らせた．各表現語について実験協力者の過半数が同義と回答した表現語の集合は，最も了解性が高いものを残し，それ以外は削除した.

4.3 結果

前節にて抽出された表現語 40 語のうち，6 組 14 語が同義と判定され，これらのうち了解性の高い 6 語を残した．この 6 語と同義でない 26 語を合わせた 32 語 (表 2) が抽出された.

5 声質表現語の類似性の調査

5.1 実験協力者

大学生 18 名が調査に参加した.

5.2 刺激音

青二プロダクションの web ページ[1]にて公開されている表 3 に示す女性声優 7 名のボイスサンプル

表 2: 同義性の調査により抽出された 32 語.

落ち着いた	愛嬌のある	愛くるしい
甘えた	軽やかな	柔らかな
若々しい	癒やされる	あたたかい
神秘的な	おしとやかな	きれいな
活発な	凛とした	明るい
女性的な	優しい	おっとりした
高い	ほんわかした	甘い
幼い	はきはきした	上品な
穏やかな	静かな	心地よい
大人っぽい	生き生きとした	色気のある
可愛い	気持ちのこもった	

表 3: 聴取実験にてボイスサンプルを使用した女性声優 7 名.

石橋桃	伊藤かな恵	井上麻里奈
桑島法子	佐倉綾音	前田愛
悠木碧		

計 20 を刺激音とした．これらのボイスサンプルは，各声優がそれぞれ異なる仮想のキャラクタでそれぞれ異なる発話内容を演じたものであり，時間長は約 10 秒間である．Web 上のボイスサンプルを Wondershare DemoCreater を用いて PC に標本化周波数 44.1 kHz にて保存し，刺激音とした.

5.3 方法

刺激音をオーディオインタフェース (Roland Rubix22)，ヘッドフォン (Sony MDR-CD900ST) により提示し，表現語 32 語について 1 を「全く〜でない」，7 を「非常に〜である」とする片側 7 段階評定尺度にて評価させた．刺激音の聴き直しは許した.

各表現語に対する評価を実験協力者間で平均し，ノンパラメトリック手法の Goodman-Kruskal の順序連関係数 γ [4, 5, 6, 7] を用いて表現語間の相関を求めた．表現語 i，j の間の順序連関係数 γ_{ij} を 1 から減じた値

$$D_{ij} = 1 - \gamma_{ij} \tag{1}$$

を表現語 i，j の間の距離と見なして．クラスタ分析を行った [3]．距離速度にはユークリッド距離，結合ルールには完全連結法を用いた.

[1] www.aoni.co.jp/actress/

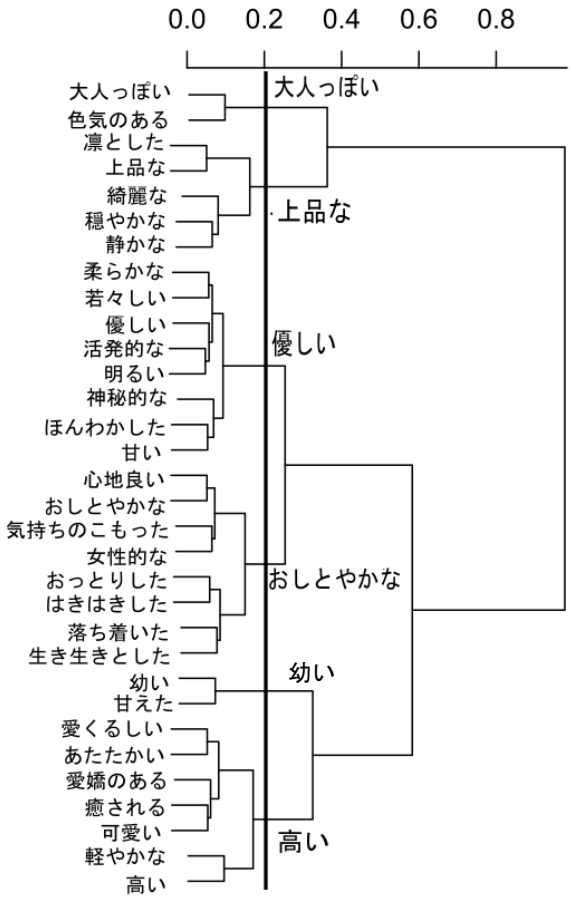


図 2: 類似性に関する聴取実験により得られたデンドログラム.

5.4 結果

クラスタ分析により表現語の分類を試みた結果のデンドログラムを図 2 に示す．分類点を 0.2 に設定し，表現語を 6 個のクラスタにまとめ，それぞれを代表する表現語でラベル付けした．6 個のクラスタには,「大人っぽい」,「上品な」,「優しい」,「おしとやかな」,「幼い」,「高い」を付与した．これらの表現語の「大人っぽい」と「幼い」を対にし，それ以外の表現語には反意語を割り当て，表 4 に示す 5 対の表現語対を決定した.

6 女性声優の声質評価の例

6.1 実験協力者

大学生 30 名が調査に参加した.

表 4: 本研究によりにより抽出された女性声優の声質表現語対.

大人っぽい声	–	幼い声
上品な声	–	荒々しい声
優しい声	–	冷たい声
元気な声	–	おしとやかな声
高い声	–	低い声

6.2 刺激音

5 節にて用いた声優 7 名のボイスサンプルから各話者 1 サンプルを選択し，そのうちの約 3 秒間の区間を切り出して刺激音とした．付録にそれらの発話内容を示す.

6.3 方法

実験はリモート環境にて実施した．実験協力者各自のヘッドフォンまたはイヤフォンにて刺激音を聴取し，表 4 の声質表現語対を両極 7 段階評定尺度にて評価させた．刺激音の聴取および回答には Google Forms を利用した.

6.4 結果

実験により得られた各声質評価語の 7 段階の評価に 0 から 6 の数値を割り当て，実験協力者間の平均値を求めた．その結果を話者ごとにプロットしたレーダーチャートを図 3 から図 9 に示す．似た声質を持つ佐倉綾音 (図 7) と悠木碧 (図 9) が似たグラフ形状を示しており，本研究にて抽出した声質表現語対には一定の信頼性があることを示唆している.

7 考察

本研究では，木戸と粕谷 [2, 3] が一般話者の声質を表現する日常語を抽出した方法論にのっとり，女性声優の演技音声の声質表現語対を抽出した．すなわち，声質表現語の収集，了解性の調査，同義性の調査，類似性の調査を通して声質表現語を絞り込み，その結果から 5 対の声質表現語対 (表 4) を得た．本研究では，声質表現語の類似性の調査の際に用いたボイスサンプルの話者が 7 名に限られている上，それぞれの声優の発話内容やそこに込

められた感情が異なる．そのため，これらの点が類似性の評価に影響している可能性が高い．

また，本研究では，クラスタ分析において分類点を 0.2 と設定し，5 対に凝縮したが (図 2 参照)，この値をシフトさせることによって表現語を増減できる．つまり，目的によって声質表現の分解能を調整することが可能である [3]．

なお，本研究にて得られた女性声優の声質表現語対と木戸と粕谷 [3] が抽出した声質表現語 (表 1) の間で重複するものは「高い声–低い声」のみである．この結果は，声優の演技音声と一般話者の音声とでは評価軸が異なることを示唆している．

本研究では，さらに得られた声質表現語対を用いて女性声優のボイスサンプルの評価を試みた．それによって得られた図 3 から図 9 の図は，女性声優の声質を表すプロファイルとして活用できる．また，これを多次元ベクトルとみなし話者間の距離を評価することも可能である．ただし，この聴取実験も刺激音の発話内容およびそこに込められた感情が声優ごとに異なり，その影響を受けていることが否定できない．

8　おわりに

本研究では女性声優の演技音声を対象にして声質表現語対を抽出する作業を行った．その結果,「大人っぽい声–幼い声」,「上品な声–荒々しい声」,「優しい声–冷たい声」,「元気な声–おしとやかな声」,「高い声–低い声」の 5 対の声質表現語対を得た．

謝辞

本研究の一部は，大学間連携等による共同研究の支援を受けた．

参考文献

[1] 声優グランプリ, 主婦の友社, 3 月号 (2022).
[2] 木戸博, 粕谷英樹, 通常発話の声質に関連した日常表現語の抽出, 日本音響学会誌, 55(6), 405–411 (1999).
[3] 木戸博, 粕谷英樹, 通常発話の声質に関連した日常表現語: 聴取評価による抽出, 日本音響学会誌, 57(5), 337–344 (2001).
[4] L. A. Goodman and W. H. Kruskal, “Measures of association for cross classifications,” *J. Am. Stat. Assoc.*, 49(268), 732–764 (1954).
[5] L. A. Goodman and W. H. Kruskal, “Measures of association for cross classifications 2: Further discussion and references,” *J. Am. Stat. Assoc.*, 54(285), 123–163 (1959).
[6] L. A. Goodman and W. H. Kruskal, “Measures of association for cross classifications 3: Approximate sampling theory,” *J. Am. Stat. Assoc.*, 58(302), 310–364 (1963).
[7] L. A. Goodman and W. H. Kruskal, “Measures of association for cross classifications 4: Samplification for asymptotic variances,” *J. Am. Stat. Assoc.*, 67(338), 415–421 (1972).

付録 6 節の声質評価実験にて用いたボイスサンプルの書き起こし

ボイスサンプルは青二プロダクションの web ページよりダウンロードしたものである．

- 石橋桃「ちょっとまいちゃん泣いちゃったじゃん.」
- 伊藤かな恵「そういうことはよくないと思います.」
- 井上麻里奈「絶対あたしのこと目の敵にしてる.」
- 桑島法子「半妖の犬夜叉って奴を殺しに行く.」
- 佐倉綾音「何選んでも素敵に着こなしちゃうのよね.」
- 前田愛「今日はお友だちのお家 (うち) に行ってきます.」
- 悠木碧「空き地の子猫ちゃんにあげに行こうよ.」

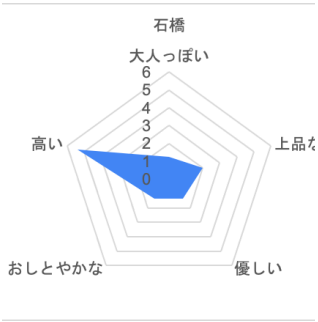


図 3: 石橋桃の評価結果.

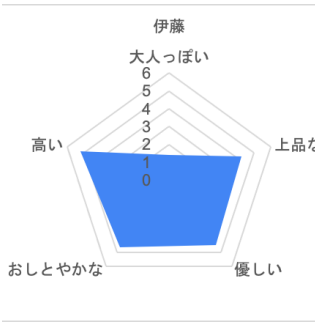


図 4: 伊藤かな恵の評価結果.

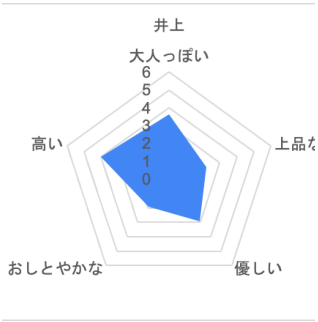


図 5: 井上麻里奈の評価結果.

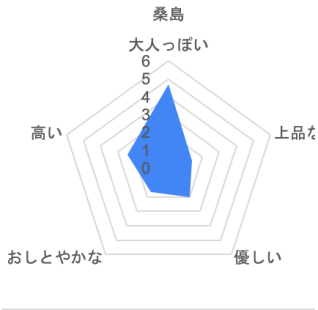


図 6: 桑島法子の評価結果.

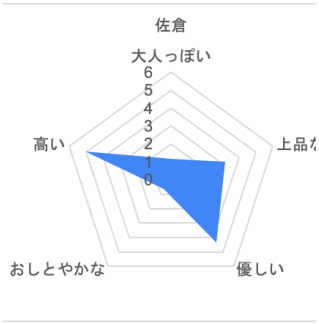


図 7: 佐倉綾音の評価結果.

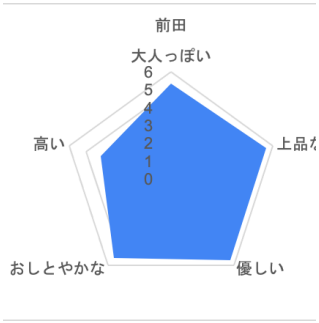


図 8: 前田愛の評価結果.

図 9: 悠木碧の評価結果.

2023年3月3日, 日本音響学会音声コミュニケーション研究会資料, Vol. 3, No.2, SC-2023-11
2023年3月3日, 日本音響学会音声研究会資料, Vol. 3, No.2, SP-2023-14

雑談対話における文脈と発話交代を考慮した応答文選択法

袖谷紳太郎 † 鈴木基之 ‡

† 大阪工業大学情報科学部 〒573-0171 大阪府枚方市北山１丁目７９－１
E-mail: †m1m21a27@oit.ac.jp, ‡moto@m.ieice.org

あらまし 近年，雑談対話システムの研究分野において，GPT を用いる方法が注目されている．この方法では，状況や対話履歴等を GPT に入力することで，応答を自動生成することができるが，必ずしも自然な応答を生成できるわけではなく，文脈と矛盾したり，対話相手の立場のような発言を生成したり，といった誤りを起こす場合がある．そこで，GPT から複数の応答候補文を生成し，この中から別のモデルを使って適切な応答文を選択する方法が提案されている．本論文では SentenceBERT を特徴抽出器として利用し，少量のデータで再学習させることで，文脈に矛盾しない事，発話者の立場をとり違えない事を判定するモデルを作成し，その性能評価を行った．
キーワード 雑談対話，応答選択，機械学習

Study on response selection method considering context and turn taking in chatting

Sodeya SHINTARO† and Motoyuki SUZUKI‡

† Osaka Insutitute Technology, 1 Chome 79-1 Kitayama, Hirakata, Osaka, 573-0018 Japan
E-mail: †m1m21a27@oit.ac.jp, ‡moto@m.ieice.org

Abstract In recent years, methods using GPT have attracted attention in the research field of chat dialogue systems. In this method, responses can be automatically generated by inputting situations and context. into GPT, but it is not always possible to generate natural responses. This may cause errors such as generating silly remarks. Therefore, methods have been proposed in which multiple response candidate sentences are generated from GPT and an appropriate response sentence is selected from among these using a selection model. In this paper, we used SentenceBERT as a feature extractor and re-learned it with a small amount of data to create a model that is consistent with the context and the speaker, and evaluated its performance.
Key words Chat dialogue, Response selection, Machine learning

1. はじめに

対話システムにおける応答生成には，手作業で作成したルールに則って対話を行うルールベース手法，質問文と応答文のペアを大量に用意してデータベースを構築し，ユーザの発話文と類似する質問文を検索し，その質問文とペアとなる応答文を発話に用いる用例ベース手法が古くから用いられて来たが，雑談対話においては対話のパターンが無限大にあり，また過去の発話の流れなども考慮しなければならないことからこれらの手法では実現が難しかった．しかし近年では，ニューラルネットワークによってその都度発話を自動生成する生成ベース手法が提案され，雑談対話システムの研究分野において注目を集めている．

自然な応答生成を実現したニューラル対話モデルとして，杉山ら [1] が提案した対話データで学習した Transformer Encoder-Decoder モデルや，山崎ら [2] の GPT に Few-Shot Learing を施すことで対話を行うモデルがある．このようなニューラル対話モデルは雑談対話の場面で自然に会話することが出来る上に，一度モデルを学習してしまえばルールベース手法や用例ベース手法のように人手によるルールの作成やデータベース構築などの作業を必要としない強みがある．しかし，モデルが生成する応答の中には，対話の流れとして不自然な応答も存在するため，何らかのフィルタリングによって生成された不自然な応答を除去する必要がある．山崎らは，「ありがとうございます.」などの情報量が少なく話が広がらない会話，「そろそろ寝なきゃいけません.」などの会話を終了してしまう会話，攻撃的，差別的な内容を含んだ会話や，URL などの雑談にそぐわない禁止語句を事前に人手で収集しておき，生成された応答との類似度を計ることによってフィルタリングを行った．この手法はニューラル対話モデルによる自然な会話を実現したが，質の高いフィル

タリングをするためにはそれぞれのフィルタに用いる数多くの不適切な会話例を人手で収集する必要がありコストがかかってしまう問題がある．

また，長澤ら [3] は，対話モデルが生成した応答候補文の中から，複数の指標を用いて最も自然な文章を選択する手法を提案した．その指標の一つに，文脈を考慮した上でシステムの応答として最も自然な文を選択する方法があった．この方法では，事前学習モデルである BERT [4] をベースとして，今までの対話履歴に対して応答候補文が適切かどうかを判定するよう，fine-tuning したモデルを構築している．しかしこのモデルでは，応答候補文のトピックが対話履歴のトピックに類似していれば，文脈に対して矛盾していたとしても高いスコアを出してしまう，という問題点があった．

そこで本研究は，BERT よりも文脈を的確に反映することができる SentenceBERT [5] を利用し，最も適切な応答文を自動で選択する方法を提案する．SentenceBERT は，BERT を文章のベクトル化に特化させる様に fine-tuning したモデルであり，文章の分類や類似度計算などに用いられている．文章のベクトル化に特化することで，文章中の文脈等がより明確に考慮されることが期待できるため，これを用いることで，より適切な応答文の選択ができるものと思われる．

2. SentenceBERT

SentenceBERT は，BERT をベースに fine-tuning したモデルで，文章を入力としてクラスタリングや推論を行う場合に BERT より高速で高性能な結果を残している．長澤らの研究では事前学習モデルに BERT を用いていたが，対話は複数の文から構成されるため，BERT よりも SentenceBERT を用いる方が性能が向上すると考えられる．

2.1 SentenceBERT によるベクトルの特徴

長澤らはトピックが類似していれば文として矛盾があったとしても高スコアを出してしまうという問題点を挙げていたが，SentenceBERT は BERT よりもトピックに左右されないベクトル化を行うことが出来る．例えば，S_1={大学構内では喫煙禁止です。}，S_2={学校でタバコを吸うのはダメです。}，S_3={今日は学校でタバコを買った。}の 3 文を SentenceBERT と BERT でそれぞれベクトル化しお互いのコサイン類似度を計算したところ，表 1 のようになった．SentenceBERT ではタバコに関してネガティブな内容である S_1，S_2 の類似度が最も高く，それに比べてポジティブな内容である S_3 との類似度は低くなった．一方で，BERT ではどの組み合わせの類似度もほとんど変わらない結果となっている．このことから SentenceBERT は BERT に比べて，文のトピックだけではなく意味や意図を考慮したベクトルを出力出来る事が分かる．そのため，文章のトピックに惑わされずに応答文を選択できる事が期待される．

また，SentenceBERT が文脈を理解できているのであれば，同じ文に対応するベクトルであっても文脈が変わることによって大きく異なったベクトルとなる事が予想される．S_4={こんにちは、今日はいい天気だね。そうだね、でも日陰は寒くて冬って感じ。いよいよってかんじだよね。もう十二月らしいよ。もう年も明けちゃうなあ。}，S_5={今年っておせち料理はどうするんだっけ。お店に頼んだよ、年越しそばは作るけど。へー。あ、除夜の鐘が聞こえてきたぞ。もう年も明けちゃうなあ。}の 2 文章のように，文末の表現を「もう年も明けちゃうなあ。」で統一した文章のベクトル化を行った後に，それぞれの「もう年も明けちゃうなあ。」の部分に相当するベクトル間でコサイン類似度を計算したところ，SentenceBERT は 0.74，BERT は 0.98 であり，BERT で計算される文章ベクトルは直前の文章の影響を受けにくいのに対し，SentenceBERT は文脈を考慮したベクトルを計算するため類似度が BERT よりも低くなることが分かった．以上の事から，応答選択の事前学習モデルとして SentenceBERT の方が有効であることが期待できる．

表 1　SentenceBERT と BERT による文章間のコサイン類似度 (右上が SentenceBERT，左下が BERT によるもの)

文章	S_1	S_2	S_3
S_1: 大学構内では喫煙禁止です。		0.66	0.37
S_2: 学校でタバコを吸うのはダメです。	0.86		0.48
S_3: 今日は学校でタバコを買った。	0.86	0.82	

3. 応答文の選択手法

本研究ではニューラル対話モデルによって複数の応答候補文の生成を行い，応答候補文の中から最も適切な応答文を選択する．応答文選択の方法として以下の 3 つの方法を提案する．文章をベクトルに変換するモデルとして，いずれの手法においても SentenceBERT [5] を利用した．

3.1 応答文の分類による手法

この手法は，長澤らが行った文脈として最も自然な文を選択する方法を踏襲したモデルである．対話履歴と応答候補文の 1 文を連結してモデルに入力し，その応答候補文が対話履歴に対して文脈として自然か不自然かを分類する．出力の最後では Softmax 関数が自然・不自然のそれぞれのラベルに対し確率を出力するため，その中の自然である確率が最も高かった候補文を応答として決定する (図 1)．

モデルの構造は図 2 のようになっており，対話履歴と応答候補文からなる入力文を SentenceBERT でベクトル化したのちに，各ベクトルを平均したベクトルを全結合層に入力して分類を行う．ここでは全結合層のみを学習する「文ベクトル分類モデル」と SentenceBERT も含めて学習を行う「Fine-tuning モデル」を提案する．モデルは分類問題を学習するため，学習時には応答候補文が対話履歴に対して文脈として自然な正例と，応答候補文が対話履歴に対して文脈として不自然な負例が必要である．

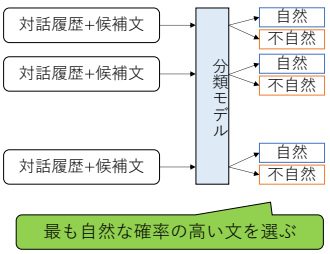


図 1　分類モデルによる応答文選択手法

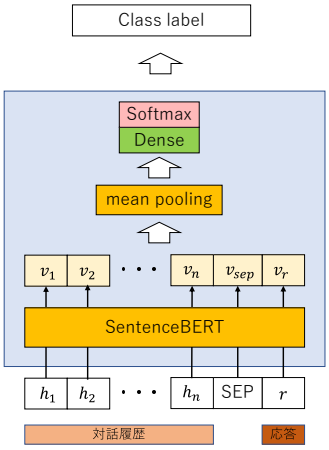


図 2　分類モデルのモデル構造

3.2　応答文の推定による手法

対話は複数の人間がターンを取り合って発話を行う時系列的なやりとりであるため，直前までの対話の履歴から次にどのような内容の発話文が続くか推測することが可能である．そこで，対話履歴に含まれる発話 $h = \{h_1, h_2, ...h_n\}$ をそれぞれ独立にベクトル化した時系列データ $v = \{v_1, v_2, ...v_n\}$ から，1 段の単方向 LSTM を用いて次に来る文のベクトル v_e を予測するモデル (図 3) を学習し，それを用いて適切な応答候補文を選択するする方法を提案する．この手法では，対話履歴から推定した次の発話のベクトルと，応答生成モデルが出力した応答候補文を SentenceBERT でベクトル化したものの間でコサイン類似度を計算し，最も類似度の高かった応答候補文を応答として決定する．この手法で用いる推定モデルを，「文ベクトル推定モデル」とする．

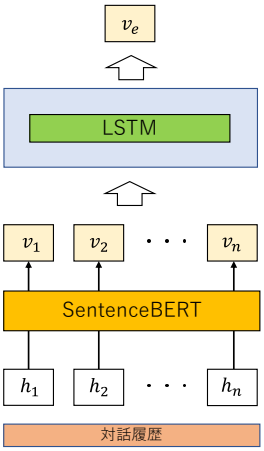


図 3　分類モデルのモデル構造

4.　応答文選択の実験

4.1　使用したコーパス

学習コーパスとして NTT が公開している JPersonaChat [6] を用いた．JPersonaChat はキャラクターの設定を与えられた 2 人が，それに基づいて雑談を行う対話を 5000 対話収録している．JPersonaChat には対話データとペルソナデータが含まれるが，本研究では対話データのみを用いてこの内 4,000 対話を訓練データ，500 対話を検証用のデータとして学習を行い，残りの 500 発話を用いて評価を行った．また，評価のためのコーパスとして STUDIES [7] を用いた．STUDIES は「個別指導塾の女性講師が，勉強の合間に生徒と雑談している」シチュエーションを想定した対話コーパスであり，10–20 ターンで終わる対話が 150 台本，4 ターンで終わる対話が 720 台本収録されている．本研究では，10 ターン以上の対話のみを用いた．これらのコーパスには，文ベクトル分類モデルと Fine-tuning モデルの学習を行うための負例が存在していない．そのため，コーパス内の別の対話からランダムに抽出することによって負例を生成した．負例は訓練データと検証用のデータには 1 文，評価データには 3 文与えた．

しかし，負例をランダムで与えた場合，いくつかの問題が考えられる．1 つは負例が対話履歴とは全く関係のない文章が与えられる可能性があることである．この様な負例は，文脈が自然かどうかに関わらずトピックの違いだけで応答文として正しいか判断で来てしまう．また，逆に負例であるにもかかわらず文脈的に成立してしまい，負例としては不適切な文章が負例に選ばれてしまう可能性もある．

佐藤ら [8] はこのような問題に対処するために，負例を厳選したコーパスの構築を行った．佐藤らは正例と類似する文章をデータベースから複数文抽出し，その後人手の評価で負例として不適切であると思われる文を除去する事で，対話履歴に対してトピックは似ているものの文脈として成立しない負例を作成した．このコーパスは英語で構築されているため，DeepL [9] で和訳した後に人手で不自然な翻訳がされている文章は除去し，評価データとして用いた．

4.2　モデルの評価

JPersonaChat，STUDIES，佐藤らのデータを用いて，提案した 3 つのモデルの評価を行った．評価の方法として，7 発話からなる対話履歴に対して 1 文の正例と 3 文の負例を与えた 4 文のセットを作り，正例を選べるかどうかで評価した．また，SentenceBERT を用いていないモデルと性能を比較するために，長澤らが継続度スコアの算出に用いたモデルとの比較も行った．表 2 に結果を示す．JPersonaChat で評価した場合においては，Fine-tuning モデルが 92.9% と最も良い性能を示したが，STUDIES では 72.0%，佐藤らのデータでは 69.0% を文ベクトル推定モデルが示し，最も良い結果になった．Fine-tuning モデルと長澤らの結果は各コーパスにおいて大きく差が開かなかったことから，事前学習を BERT から SentenceBERT に変えた効果は確認できなかった．

文ベクトル推定モデルとそれ以外の 3 つのモデルを比べると，文ベクトル推定モデル以外のモデルは STUDIES・佐藤らのデータで評価した時に性能が大きく低下していることが分かる．これらのモデルは文ベクトル推定モデルに比べて，負例のトピックが正例・対話履歴のトピックと類似している対話例 (表 4) を間違えやすい傾向があった．対話コーパスの特性として，STUDIES は JPersonaChat に比べて収録されたどの対話においても学校での話題が頻出するため，ランダムで選択した負例に

表 2　提案したモデルの正解率

	JPersonaChat	STUDIES	佐藤らのデータ
文ベクトル推定モデル	86.7(%)	72.0(%)	69.0(%)
文ベクトル分類モデル	89.6(%)	58.2(%)	40.9(%)
Fine-tuning モデル	92.9(%)	66.0(%)	47.1(%)
長澤らのモデル	92.7(%)	69.1(%)	45.4(%)

「テスト」等の対話履歴や正例とトピックが類似した文が含まれやすい傾向がある．学習に用いた JPersonaChat ではこのような負例が与えられる事が少ないため，対話履歴と応答候補文にトピックの繋がりがあるかどうかで正例と負例の分類を学習してしまい，STUDIES・佐藤らのデータで評価した時の性能低下に繋がったと考えられる．また，これらのモデルは正例が「ありがとう」などの汎用的でトピックのない対話例 (表 5) に関しても選択を間違える事が多かった．このことからも，文ベクトル推定モデル以外の 3 つのモデルは対話履歴と応答候補文のトピックの繋がりを重視して応答文選択を選択していると考える事が出来る．一方で，文ベクトル推定モデルではこのような間違いを抑制する事が出来た．

また，全モデルに共通して正解できなかった対話例に，正例が話題転換を行っている対話例 (表 6) があった．これは対話履歴と正例のトピックが全く異なる為に選択できないのだと考えられる．また，負例が最後の文と同一話者が続けて発話しているように思える対話例 (表 7) も選択を間違える事が多かった．直前の発話者が A であった時に，B の発話としては文脈的に不自然だが，A が更に続けて発話したと考えれば文脈的に自然な文となっている．応答を生成するモデルは入力の文章の続きを生成するモデルを対話用に学習している事が多いため，このような応答は十分に考えられる．

5.　応答選択における話者推定

雑談対話で用いられる応答文生成モデルは，もともと入力に対して続きの文を生成するモデルであるため，最後の文と同一話者による続きの発話を応答候補文として生成してしまう可能性は十分に考えられる．このような文を選択しないようにするために，応答候補文が文脈として自然かどうかの判定に加えて，候補文の話者がユーザとシステムのどちららしいかを併せて推定する事によってこの問題を解決することを提案する．

応答候補文の選択手法は 3.1 節と同一の手法を用い，モデル構造も Fine-tuning モデルと同じだが入力文が異なる．図 4 のように対話履歴に含まれる各発話文の直前に，その発話文がどちらの話者によるものかの情報を与えた文を入力することで，最後の文と同一話者が続けて発話しているように思える文が与えられた場合に発話者が B であるにも関わらず発話内容が A であるという文がモデルに与えられることになり，その差異によって正しい応答選択が出来るようになる事を期待する．

Fine-tuning モデルと話者情報を考慮したモデルを JPersonaChat，STUDIES，佐藤らのデータを用いて評価し比較したところ，性能は表 3 のようになった．話者情報を考慮したモデルは Fine-tuning モデルに比べ JPersonaChat では 0.2 ポイント性能が低下したが，STUDIES では 4.4 ポイント，佐藤らのデータでは 1.4 ポイントの性能向上が見られた．話者情報を考慮したモデルは Fine-tuning モデルに比べて，問題だった負例が最後の文と同一話者が続けて発話しているように思える対話例 (表 7) で幾つかの例を正解する事に成功し，性能向上に繋がったと考えられる．また，正例が「ありがとう」などの汎用的でトピックのない対話例 (表 5) に関しても話者情報を考慮した方が正しい選択を出来る傾向があった．しかし対話履歴・正例と負例のトピックが類似している場合には文脈として不自然でも負例を選択してしまう傾向は大きくは改善されなかった．

表 3　話者情報を考慮したモデルの正解率

	JPersonaChat	STUDIES	佐藤らのデータ
話者情報を考慮したモデル	92.7(%)	70.4(%)	48.5(%)
Fine-tuning モデル	92.9(%)	66.0(%)	47.1(%)

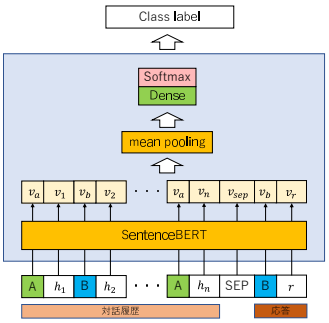


図 4　話者情報を考慮したモデルの構造

6.　ま　と　め

本稿では，ニューラル対話モデルが生成した複数の応答候補文の中から最適な応答文を決定する手法として，SentenceBERT を利用し，対話履歴に対し文脈として最も自然な応答候補文を選択する手法と，対話履歴の次に来る文を予測しその文と応答候補文の類似度から選択する手法を提案し，実験・評価を行った．結果としては，学習に用いていない評価データに対して文ベクトル推定モデルが STUDEIS に 72.0%，佐藤らのデータに 69.0% と最も良い性能を示した．しかし，正例が話題転換を行っている対話例や負例が最後の文と同一話者が続けて発話しているように思える対話例に関しては正しく選択できなかった．負例が最後の文と同一話者が続けて発話しているように思える対話例では選択を間違えてしまう問題への対処として話者情報を考慮した応答選択を提案し，実験・検証を行った結果，話者情報を考慮することによって Fine-tuning モデルと比べ STUDIES では 4.4 ポイント，佐藤らのデータでは 1.4 ポイントの性能向上が見られたが，文ベクトル推定モデルの性能を超える事は出来なかった．

文　　献

[1] 杉山弘晃　他. Transformer encoder-decoder モデルによる趣味雑談システムの構築．人工知能学会研究会資料言語・音声理解と対話処理研究会 90 回，p. 24．一般社団法人人工知能学会，2020
[2] 山崎天　他: ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム，人工知能学会研究会，pp113 (2022)

[3] 長澤春希　他. aoba v2 bot: 多様な応答生成モジュールを統合した雑談対話システム. 人工知能学会研究会資料言語・音声理解と対話処理研究会 93 回，p．101. 一般社団法人人工知能学会，2021
[4] Jacob Devlin et al．BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding．arXiv:1810.04805 [cs.CL], 2018
[5] Nils Reimers，Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP 2019, p．3982. Association for Computational Linguistics，2019.
[6] https://github.com/nttcslab/japanese-dialogtransformers
[7] 齋藤佑樹　他:“STUDIES：表現豊かな音声合成に向けた日本語共感的対話音声コーパス,” 日本音響学会 2022 年春季研究発表会講演論文集，2-3P-15，pp．1133–1136，2022 年 3 月．
[8] 佐藤志貴　他: 負例を厳選した対話応答選択による対話応答生成システムの評価，自然言語処理，vol．29 No．1，2022 年 5 月
[9] DeepL，https://www.deepl.com/translator

表 4　負例のトピックが対話履歴と類似している例

対話履歴
話者 A: わかります？物理のテスト、すっごく良かったんです！
話者 B: おめでとう！難しいって言いながら、頑張っていたから。
話者 A: それもあるけど、先生が「出るよ」って言ったところ、全部出たんですよ！びっくり！
話者 B: えっ、ほんと？それは良かった。あのリスト、役に立ったのね。
話者 A: 見てくださいよ。９０点です！テスト返してもらう時、先生にも「頑張ったな！」って言われました！
話者 B: それはすごいね！
話者 A: 先生の、テストの予想のほうが、すごいですよ！
応答候補文
正例: あの先生は私も学生の頃に習ったから、なんとなく傾向がわかるの。
負例: 中学の時に、県大会へ行ったことはないから、ちょっと不安だけどね！
負例: 先生、聞いてくださいよ！俺、納得いかないことがあるんです！
負例: このあいだのテストでは、かなりいい点数が取れました。

表 5　正例が短く情報が少ない例

対話履歴
話者 A: わたしは両親と一緒に暮らしてます。一人暮らししてると、親御さん、心配しません？
話者 B: はい、とても心配です。わたしがこんな性格だからかもしれませんが
話者 A: いやいや、そんな悲観しないでください。考え方を変えれば、慎重だということですよ！
話者 B: ありがとうございます、さすが僧侶さんということでお優しいですね。
話者 A: いやいや、一生修行です！わたしなんて、まだ 20 代そこそこの、ひよっこですから！精進あるのみです。
話者 B: そうですね、わたしもあなたを見習ってパティシエの実績を残します！ありがとうございます！
話者 A: いえいえ、頑張ってくださいね！もし悩むことがあれば、気軽にお寺に相談に来てください！
応答候補文
正例: はい！
負例: ありがとうございます！いつか！仕事で世界中を飛び回れるように、頑張ります！
負例: 不満はないのですが、このご時世で飲食店も壊滅的で違う業種に挑戦してみたいという考えです。
負例: それはいいね。４コマ漫画の面白さを知ってるあなたは上級者だよ。わたし実は先日、ゲーム制作会社に就職が決まったんだ。ゲームはやる？

表 6　正例が話題転換をしている例

対話履歴
話者 A: おお。兄弟が多い人あるあるだね。わたしは一人っ子だったから、丸まる 1 本、いや、2 本は食べてたな。
話者 B: おお、いいですね。それが、今はシングルファーザー。だから焼き芋は今度は子供と分け分けなんです。
話者 A: なんだか、良い話だなぁ。わたし、先日ゲーム制作会社に就職したんだけど、ゲームにそんなハートウォーミングな話追加したいくらいだよ。
話者 B: はは。そんなストーリーがあったら、わたしがモデルと皆にえばれるなあ。ちなみに尊敬する人はエジソンです。それもストーリーに入れてくださいね。
話者 A: あはは、わかったよ。わたし、アニメオタクで、学生時代は漫研に入っていたんだけど、あなたはアニメって、興味ある？
話者 B: アニメですか。うーん、人気なものは観てたけどそれくらいかな。わたしは運動部出身なので。
話者 A: そうなのか。残念だな。アニメについて語りたかったんだけど。
応答候補文
正例: すみません。車のことなら語れるんですが。今自動車整備士をしているもので。
負例: ええ、お聞かせしますよ。一緒に食事でもしながら話しましょう。
負例: あらそうなのか、まあいいじゃないか。わたしだって散歩が好きな 40 代だしな
負例: 経理の仕事をしているんですが、責任ある仕事を任されるようになりました。昔から数学を勉強していて、数字に強いのが評価されました。お住いはどちらなんですか。

表 7　候補文の負例が直前の文章の続きであれば自然な例

対話履歴
話者 A: 私は怒りっぽいところはあるけど、そういうのは大丈夫。尊敬している母親を見習ってるから。
話者 B: よかった。わたし声が大きいけど起こったことなもんね。お母様に感謝だな。
話者 A: そう。あなたはご家族はどうしてるの？
話者 B: わたしは両親と一人っ子のわたしの核家族。ちなみに暑い地域で生まれ育ったよ。だからここの寒さが辛くて辛くて。
話者 A: 私も沖縄に住んでいたことがあるんだ。だからあの暖かい海が懐かしくて。だから今も海の側に住んでる。
話者 B: 海に近いところなんて羨ましいな！今度遊びに行かせてもらってもいい？
話者 A: もちろん、歓迎するよ！
応答候補文
正例: ありがとう！手土産は何がいい？
負例: 良かったら今度お休みにでも遊びに来てください。お勧めのお店を案内しますよ
負例: いいですね。今度またお話聞かせてください。
負例: あなたもバナナが大好きとか？申年だけに。

2023年3月3日, 日本音響学会音声コミュニケーション研究会資料, Vol. 3, No.2, SC-2023-12
2023年3月3日, 日本音響学会音声研究会資料, Vol. 3, No.2, SP-2023-15

唇動画像からの音声生成法における入力特徴量の単純化に関する検討

金澤 尚希† 鈴木 基之†

†大阪工業大学大学院 〒573-0196 大阪府枚方市北山 1-79-1
E-mail: †m1m21a14@oit.ac.jp

あらまし 近年，唇動画像からの音声生成を行う研究は数多く行われている．従来の手法では CNN や RNN を用いた DNN モデルで音声波形の生成を行っているものが多い．その場合，入力となる唇動画像には肌の色やホクロといった話者固有の特徴を学んでしまい，学習話者以外の話者のデータを入力とする場合には性能が低下してしまう．そこで，学習話者以外の話者においても高い性能で音声波形を生成するために，入力特徴量から話者固有の特徴を取り除く手法を提案した．本稿では提案した入力特徴量を用いて音声波形を生成し，それらを客観評価値 STOI を用いて評価した．結果として，提案した手法では唇動画像を入力した場合に比べ性能は劣化したが，話者の違いによる劣化を抑えることができる効果を確認した．

キーワード サイレント音声認識

A Study on Simplification of Input Features in Speech Generation Method from Lip Video Images

Naoki KANAZAWA† and Motoyuki SUZUKI‡

†Osaka Institute of Technology 1-79-1 Kitayama, Hirakata-shi, Osaka, 573-0196 Japan
E-mail: †m1m21a14@oit.ac.jp ‡moto@m.ieice.org

Abstract In recent years, there have been several studies on speech generation from lip video images. Many conventional methods use DNN models based on CNNs or RNNs to generate speech waveforms. In such methods, the model learns speaker-specific features such as skin color and moles, and the performance degrades when data from speakers other than the training speaker is used as input. Therefore, we proposed a method to remove speaker-specific features from the input features in order to generate speech waveforms with high performance for any speaker. In this paper, we generated speech waveforms using the proposed input features and evaluated them using any STOI. As a result, the performance of the proposed method was worse than that of the lip video input method, but we confirmed the effectiveness of the proposed method in suppressing the degradation caused by differences in speakers.

Keywords Silent Speech Recognition

1. はじめに

近年，唇動画像から発話内容を推定するサイレント音声認識に関する研究（例えば[1]）が数多く行われているこのような方法においては，唇動画像をそのまま入力しているが，入力画像系列には肌の色やホクロ，しわなどの発話に関係のないと思われる情報を多量に含んでいると考えられる．このような情報は話者によって異なるため，学習話者に比べそれ以外の話者では性能が大きく劣化することが予測される．

このような問題点を解決するために一度学習した DNN モデルを対象の少量の話者のデータを用いて再学習を行うことでその対象話者に対する性能を向上させる研究[2]も存在する．しかしその場合，対象話者ごとにモデルを再学習する必要があり効率的ではない．

そこで本研究では，こうした入力特徴量から発話に関係ないと思われる情報を排除したものを検討し，その性能を評価する．なお本研究では，発話内容をテキストで出力するのではなく，直接音声を生成する方法[3][4]を用いて性能評価を行った．

2. 従来の唇動画像からの音声生成法

従来から唇動画像から音声波形を生成する手法は提案されてきた。例えば文献[3]では，LipNet[1]の構造を利用して，音声波形を生成している．本来 LipNet は唇動画像から単語列を出力するものであるが，その出力をスペクトル特徴量とすることで，音声波形を生成する．具体的な生成の流れは図 1 のようになっている．音声波形を基本周波数，メルケプストラム，非周期性指標に分解し，その中の基本周波数とメルケプストラ

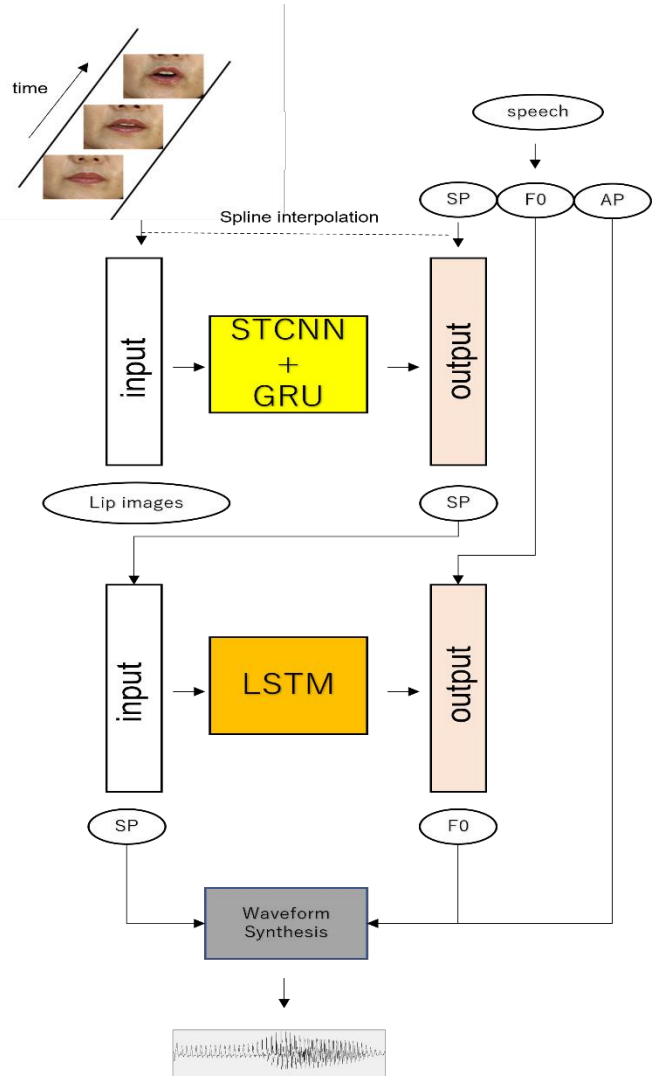


図 1 従来の唇動画像からの音声生成法

ムを DNN モデルを用いて推定する。そして，本来の非周期性指標と合わせて音声波形を生成する．メルケプストラムを推定する DNN モデルは時間方向を含んだ 3 次元の畳み込み STCNN[5]と双方向の GRU を組み合わせることで構成されており，文単位の発話を認識するのに頑強なネットワークになっている．また，これらのネットワークの入力と教師データは唇画像系列とメルケプストラムになっているが，メルケプストラムのほうがフレームレートが高いため，それに合わせて唇画像系列をスプライン補間しフレームレートを合わせることでパラレルなデータとして扱っている．基本周波数を推定する DNN モデルは LSTM を用いたものになっており，入力には推定したメルケプストラムが用いられる．このシステムでは唇画像系列からメルケプストラム，メルケプストラムから基本周波数というように連鎖的に推定を行っている．そのため，前段階の推定結果が後段に影響するという特徴がある．このシステムではスペクトルから基本周波数を推定する際，正解のスペクトルと推定されたスペクトルから推定された基本周波数は RMSE で約 12 ポイント程度の差があり，スペクトルの推定方法の改善が必要だと考えられる．

3. 入力特徴量に注目した唇動画像からの音声生成法

3.1. 入力特徴量

従来から用いられている入力特徴量は唇動画像をそのまま入力とするものである．具体的には撮影した画像から図 2 のように唇周囲を抽出したものを利用している．この特徴量は発話内容に関係のない情報を多量に含んでいると考えられる．通常，CNN を用いる場合には，CNN によって不要な情報は取り除かれるように学習されるはずではあるが，それが必ずしも発話に必要な情報のみを残したものではないと考えられるため，あらかじめこの入力特徴量より不要な情報を取り除いた入力特徴量を提案する．

発話内容を推定する上において必要とされる情報は唇の動きや形状であると考えられる．そこで唇画像から，図 3 のように顔ランドマーク検出[6]によって検出される唇の輪郭の特徴点の座標値を利用する．このようにすることで唇の形状以外の情報をすべて排除することができる．また，それぞれの特徴点を繋ぐことで図 4 のように唇の形状を単純な図形で表現したものも入力特徴量として提案する．このようにすることで特徴点を直接入力する場合と異なり，CNN を用いてより良い特徴量抽出を行うことができると考えられる．

図 2 抽出した唇画像

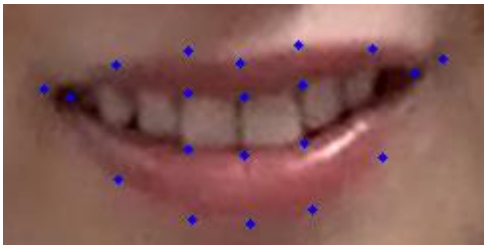


図 3 抽出された特徴点

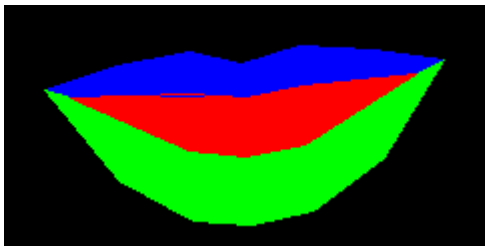


図 4　単純な図形で表現された唇

3.2. 音声生成システムの概要

唇動画像から音声波形を生成するシステムの概要を図 5 に示す．このシステムではまず，図 1 の GRU を双方向と順方向の LSTM に変更したネットワークによって唇画像系列からメルケプストラムを推定する．また，この際の入力と教師データの唇画像系列とメルケプストラムではメルケプストラムのフレームレートのほうが高いため，唇画像系列のフレームレートをメルケプストラムに合わせて補間している．

次に図 1 と同様に LSTM で構成されたネットワークを用いて推定したメルケプストラムから基本周波数の推定を行う．同時に別の LSTM で構成されたネットワークを用いて有声区間と無声区間の判定を行う．そして，推定された基本周波数の内，有声区間と無声区間の判定で無声区間と判定された区間の値を 0 にする．このようにすることで無声区間であるのに基本周波数の値が大きい場合などの修正を行うことができる．

これらの推定されたメルケプストラムと基本周波数,非周期性指標の 3 つを用いて音声波形を生成する．非周期性指標は図 1 のようには正解データのものでなくとも固定値 0 を与えれば生成される音声に大きな違いはないため固定値 0 を与えている．

入力特徴量が単純な図形で唇を表現した場合には，同様のネットワーク構造を用い，入力を唇画像系列から単純な図形で唇を表現したものに置き換えたもので音声波形を生成する．一方，入力特徴量が唇の輪郭の特徴点の場合，ほかの 2 つの入力特徴量の場合と異なり，メルケプストラムの推定を行うネットワークを変更する．唇画像系列や単純な図形で唇を表現したものでは，画像が入力となるため CNN を用いたものとなっていたが，特徴点を入力特徴量とする場合には，双方向の LSTM と順方向の LSTM で構成されたネットワークを用いてメルケプストラムの推定を行う．

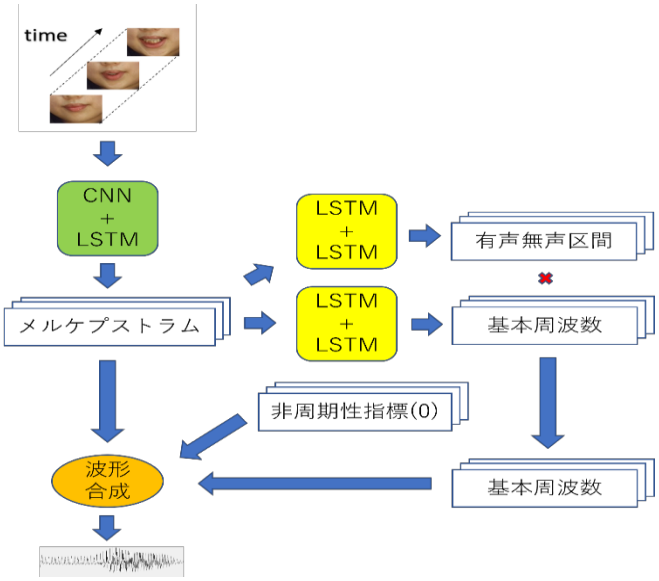


図 5　音声生成システムの概要

4. 音声の生成実験

4.1. 実験の概要

提案する入力特徴量の有効性を確認するため，音声の生成実験を行った．入力特徴量は唇画像系列，唇の輪郭の特徴点，単純な図形で唇を表現したものの 3 つを用いる．生成された音声波形は客観評価値の 1 つである STOI[7]を用いて評価を行う．この指標は劣化した音声の了解度を測る指標の 1 つであり，主観評価と相関が高いことが知られている．

音声を生成するモデルは 1 人の話者で学習した特定話者モデルと複数人の話者で学習する不特定話者モデルの 2 種類を用いる．

使用するデータベースは「ROHAN4600 マルチモーダルデータベース」[8]と「ITA コーパスマルチモーダルデータベース」[9]の 2 つである．それぞれのモデルの学習に用いたデータは表 1 のようになっている．不特定話者モデルの学習に使用するデータはそれぞれの話者が同じ内容を発話している．

表 1 学習データ数

モデル		特定話者	不特定話者
データベース		ROHAN4600	ITA
発話数	女性	1 名	3 名
合計発話数		4000 発話	900 発話
1 人あたりの発話数		4000 発話	300 発話

4.2. 特定話者モデルの性能評価

1 名の話者によるデータでモデルの学習と評価を行った結果を表 2 に示す．この結果から唇画像系列を入力特徴量とした場合が最もいい性能を示していること

がわかった．これは学習話者とテスト話者が同一であるために，肌の色やホクロといった発話に関係しないと思われる情報がいくら含まれていても常に同じであるため，性能の劣化にはつながっていないと考えられる．

一方で特徴点の座標値や単純な図形で表現したものはどちらも同程度の性能であり，唇画像系列より0.05ポイント以上低い結果になっていた．このことから提案した入力特徴量では不要な情報を取り除くのと同時に発話に関する重要な情報の一部が欠落してしまい，性能が劣化してしまっていることが分かった．

表 2 特定話者モデルの STOI 評価平均

入力特徴量	STOI
唇画像系列	0.496
特徴点の座標値	0.441
単純な図形表現	0.431

次に未知話者に対する性能を評価するため非学習話者3名によるデータでモデルの評価を行った結果を表 3 に示す．表 2 と表 3 の結果を比較すると，未知話者の場合にすべての入力特徴量で性能が劣化していることがわかる．具体的な劣化度合いは唇画像系列，特徴点の座標値，単純な図形表現がそれぞれ，33%，35%，16%程度となっていた．また，既知話者の場合と異なり単純な図形表現の場合がどの話者においても最もいい性能になっていることが分かった．これは学習話者固有の肌の色やホクロといった特徴が非学習話者と異なることが原因だと考えられる．そのため，それらを取り除いた入力特徴量である単純な図形表現が唇画像系列の性能を上回る結果になったと考えられる．

一方で特徴量の座標値を入力とした場合は他の入力特徴量に比べ大きく劣っていることがわかる．これは他の入力特徴量の場合と異なり，CNN を用いない単純なモデルであったためだと考えられる．これらのことから入力特徴量から話者固有の発話に関係のない情報を取り除くことで，話者の違いによる劣化を抑えることができると分かった．

表 3 非学習話者による STOI 評価平均

入力特徴量	話者 A	話者 B	話者 C	平均
唇画像系列	0.286	0.362	0.345	0.331
特徴点の座標値	0.266	0.297	0.297	0.287
単純な図形表現	0.345	0.369	0.367	0.360

4.3. 不特定話者モデルの性能評価

3 名の話者によるデータでモデルの学習と評価を行なった結果を表 4 に示す．評価においては学習に使用した既知話者に加え，学習に使用していない未知話者に対する評価も行った．また，未知話者に関しては学習に使用した発話内容と同様のものについての評価も行った．

表 2 と表 4 の結果を比較すると学習話者が 1 名であっても 3 名であっても既知話者の未知発話の性能は唇画像系列が最も高く，特徴点の座標値と単純な図形表現はどちらも同程度で唇画像系列の場合より劣る傾向にあることがわかる．このことから 3 名で学習を行った場合には，それぞれの学習話者の特徴を学ぶことができていたため唇画像系列の性能が最も良くなったと考えられる．

次に表 4 の未知話者についてみてみると，既知発話より未知発話のほうが高い性能になっていることがわかる．通常，既知のほうが高くなると考えられるがそうではないため，発話内容についてうまく学習ができていない可能性があると考えられる．

一方，未知発話についてみると，既知話者から未知話者に変わった際の性能の劣化度合いは唇画像系列で24%程度，特徴点の座標値で 17%程度，単純な図形での表現で 17%程度となっている．このことから，入力特徴量から話者固有の発話に関係のない情報を取り除くことで，話者の違いによる劣化を抑えることができているとわかる．

また，表 4 の結果から，どの場合においても唇画像系列を入力としたものが最も良い性能になっていることがわかった．これは表 3 の場合と異なる結果であるが，唇画像系列より単純な図形表現を入力とした場合のほうが劣化度合いは小さくなる点は同様である．しかし，特定話者モデルから不特定話者モデルになったことにより，学習話者に対する依存が少なくなったため，唇画像系列での劣化度合いは特定話者モデルの場合より 9 ポイント程度低くなっている．その結果，性能面においては唇画像系列が単純な図形表現の場合を上回ったと考えられる．これらのことから，提案する入力特徴量では発話内容を推定するには情報が不足していることがわかる．

表 4 不特定話者モデルの STOI 評価平均

評価話者	既知	未知	
発話内容	未知		既知
唇画像系列	0.471	0.357	0.316
特徴点の座標値	0.384	0.317	0.271
単純な図形表現	0.384	0.320	0.280

5. おわりに

唇動画像を入力として，音声波形を生成するシステムにおいて，入力特徴量を変更することで性能を向上させる手法を提案した．唇動画像を入力とする場合に

は，発話に関係しない情報が多量に含まれていると考えられたため，それらを取り除くことで性能の向上を目指した．具体的には唇の動きや形状のみを残すため唇の輪郭を表す特徴点とそれを用いて単純な図形で唇を表現したものの 2 つを提案した．

これらの入力特徴量を用いて音声生成を行いその音声を用いて性能評価を行った．1 人の話者のデータで学習をした特定話者モデルと複数の話者のデータで学習した不特定話者モデルのそれぞれについて実験を行った．

特定話者モデルで生成された音声について学習話者で評価を行ったところ，唇画像系列を入力特徴量としたものが最も良い性能となり，特徴点の座標値を入力特徴量としたものと単純な図形で表現したものを入力特徴量とした場合はそれぞれ同程度の性能となり，唇画像系列より 0.05 ポイント以上下回る結果となった．これらのことより，提案した入力特徴量は不要な情報を取り除くと同時に発話に関する重要な情報の一部も欠落していることが分かった．

一方，非学習話者 3 名で評価を行ったところ，単純な図形で表現したものを入力特徴量としたものがどの話者においても最も良い性能となった．一方，特徴点の座標値を入力特徴量とした場合の性能は他の入力特徴量に比べ大きく低下していた．これらのことより，入力特徴量から不要な情報を取り除くことで話者の違いによる劣化を抑えることができたが，CNN を用いない単純なモデルでは性能は大きく劣化するということが分かった．

不特定話者モデルでは，特定話者モデルの学習話者評価と同様に唇画像系列が最もいい性能であったが，未知発話について既知話者から未知話者に変わった際の性能の劣化度合いが唇画像系列を入力特徴量とした場合に比べて，特徴点の座標値や単純な図形で表現したものの場合では 7 ポイント程度劣化を抑えることができていることが分かった．このことより，提案する手法では適切に話者の不要な情報を排除することができていると分かった．

これらのことより，提案した入力特徴量より発話に関する情報を多く残した入力特徴量を模索する必要があると考えられる．

謝辞

本研究の一部は JSPS 科研費(22K12916)の助成を受けて行われた。

文　献

[1] Y. Assaeletal., “LipNet : end-to-end sentence-level lipreading,” arXiv:1611.01599, (2016)

[2] 金澤 尚希 , 鈴木 基之,``唇画像からの音声生成における話者依存性の分析''研究報告音楽情報科学（MUS）,2021-MUS-131(19),1-6 (2021-06-11) , 2188-8752

[3] 伊藤大貴，滝口哲也，有木康雄,”LipNet 構造を用いた唇画像からの音声への変換” 日本音響学会 2018 年春季研究発表会講演論文集, 2-Q-30, pp. 347-350, 2018-03.

[4] Z. He, C. -Y. Chow and J. -D. Zhang, "STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction," 2019 20th IEEE International Conference on Mobile Data Management (MDM), Hong Kong, China, 2019, pp. 226-233, doi: 10.1109/MDM.2019.00-53.

[5] Prajwal Renukanand, Rudrabha Mukhopadhyay, Vinay Namboodiri and C.V. Jawahar“Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis”CVPR, (2020)

[6] Kazemi, V. & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE conference on computer vision and pattern recognition (p./pp. 1867--1874),

[7] C. Taal et al., “Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech”, in Proc. ICASSP,pp.4214-4217, (2010)

[8] 森勢将雅,“ROHAN : テキスト音声合成に向けたモーラバランス型日本語コーパス”日本音響学会誌,vol.79,no.1,pp.9-17, Jan. (2023).

[9] 小口純矢，金井郁也，小田恭央，齊藤剛史，森勢将雅“ITA コーパス : パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価”, 情報処理学会研究報告, vol. 2021-MUS-131, no. 31, pp. 1-6, (2021)

2023年3月3日, 日本音響学会音声コミュニケーション研究会資料, Vol. 3, No.2, SC-2023-13
2023年3月3日, 日本音響学会音声研究会資料, Vol. 3, No.2, SP-2023-16

国語の教材文の初読方法

加藤 凪咲† 白勢 彩子†

†東京学芸大学 〒184-8501 東京都小金井市貫井北町 4-1-1

E-mail: †shirose@gakugei.ac.jp

あらまし 文章を読む行為の中でも特に「初読」は、文章との最初で最後の出会いであり、単元の最初で文章をどのように読むのかという初読方法は特にその効果や目的について考えることが求められると考えられる。そこで、本研究では、初読に焦点を当て、音読、黙読のうち、実際の教育現場においてどのような初読方法が一般的であるのか、また、初読方法をどのように選択・判断しているのかについて調査することを目的として、教科書の教師用指導書の調査と、国語科教員に対するアンケート調査を実施して、結果を報告する。

キーワード 国語科教育、初読、音読、教師用指導書、アンケート

Methods of initial reading of Japanese language textbooks

Nagisa KATO† and Ayako SHIROSE†

†Humanities and Social Sciences Division, Tokyo Gakugei University 4-1-1 Nukuikita-machi, Koganei-shi, Tokyo, 184-8501 Japan

E-mail: †shirose@gakugei.ac.jp

Abstract Among the acts of reading texts, "initial reading" is the first encounter with a text. It is assumed that the method of initial reading, which is how a text is read at the beginning of a unit, especially requires consideration of its effect and purpose. Therefore, in this study, in order to understand which method of initial reading is common among oral and silent reading in actual educational settings, and how the method of initial reading is selected and judged, we carried out a survey of textbook instruction manuals for teachers and a questionnaire survey of Japanese language teachers.

Keywords Japanese language education, initial reading, oral reading, textbooks for teachers, questionnaire survey

1. はじめに

従来、音読と黙読など文章の読み方を比較する研究は、文章の種類や対象者、研究目的をかえて多くの研究が行われてきた[1]。これらの研究により、子どもにとって音読することは読解力に影響を与え、内容理解の成績が良くなる傾向があるが、大人になると音読でも黙読でも内容理解の成績に大きな差は生じないということが示され、国語の授業に影響を与えている。

文章を読む行為の中でも特に「初読」は、文章との最初の出会いであり、単元の最初で文章をどのように読むのかという初読方法は特にその効果や目的について考えることが求められると考えられる。そこで、本研究では、初読に焦点を当て、音読、黙読のうち、実際の教育現場においてどのような初読方法が一般的であるのか、また、初読方法をどのように選択・判断しているのかについて調査することを目的として、教科書の教師用指導書（以下、単に「指導書」）の調査と、国語科教員に対するアンケート調査を実施して、結果を報告する。

2. 初読とは

「初読」という単語は様々な指導書に登場するが、『日本国語大辞典』（第 2 版）を含むほとんどの辞書に「初読」の項目がない。足立[2]が「初読とは、初めて読むことであるから、読んできたところまでは何が書いてあるかを理解しているが、その後に何が書かれているかは想像するしかない」と述べており、本稿では「初読」を「初めて読むこと」と定義する。

「初読」と似た言葉に「通読」がある。『日本国語大辞典』（第 2 版）に「ひととおり読むこと。始めから終わりまで読み通すこと。」とあり、本研究では「通読」を「文章を始めから終わりまで読み通すこと」と定義する。「初読」とは読み通すという点で違いがある。

他、本稿では主に「音読」「黙読」「範読」「朗読 CD」の 4 つを取り上げる。本稿では、『日本国語大辞典』[3]の記述に基づき、本稿では「音読」は「文章を声に出して読むこと」、「黙読」は「声に出さずに目だけで読むこと」、「範読」は「生徒も文章を見ている状態で教師が模範として読んで聞かせること。」と定義し、それぞれの読み方については細分化しないとする。

「朗読 CD」については教科書会社の朗読 CD に付随する文章を調査した。

3. 指導書の調査

教科書の指導書は、ある特定の学級や学校を想定して作成しているわけではなく、各教科書会社がある程度汎用性のある授業の流れや言語活動を設定していると考えられる。教育現場の平均となる考え方を見ることができると考えたことから、指導書の調査を実施した。また、音読と黙読の効果の違いがはっきりと示されていない年齢について調べる目的から、高等学校の指導書を対象とした。

3.1. 調査方法

高等学校の国語の教科書の指導書における「評論文」「小説」「随想」の教材で調査を行った。学校現場で使用されることの多い教科書を調査するため、調査対象の指導書は日本出版労働組合連合会の「資料　2019 年度高等学校教科書の採択データ」[4]における国語各教科の上位 2 社とし、全 13 種の指導書で調査を行った。調査対象を表 1 に示す。指導書の詳細な情報は、付録として文末に示す。

指導書の各教材において、以下の 6 点について調査した。

①教材名
②作者
③教材の種類
④初読で通読を行う場合にどんな方法で行うか（初読方法）
⑤初読時の指導上の留意点や初読後の活動
⑥初読の感想を求めるのか

表 1. 調査対象の指導書一覧

No.	教科名	発行者	教科書名
t1	国語総合	東京書籍	新編国語総合
t2	国語総合	東京書籍	精選国語総合
t3	国語総合	東京書籍	国語総合（現代文編）
d4	国語総合	第一学習社	改訂版新訂国語総合現代文編
d5	国語総合	第一学習社	改訂版国語総合
d6	国語総合	第一学習社	改訂版標準国語総合
d7	国語総合	第一学習社	改訂版新編国語総合
t8	現代文 A	東京書籍	現代文 A
d9	現代文 A	第一学習社	改訂版新編現代文 A
t10	現代文 B	東京書籍	新編現代文 B
t11	現代文 B	東京書籍	精選現代文 B
d12	現代文 B	第一学習社	改訂版現代文 B
d13	現代文 B	第一学習社	改訂版標準現代文 B

3.2. 調査結果

指導書の調査の結果を東京書籍(t1･t2・t3・t8・t10・t11)と第一学習社（d4・d5・d6・d7・d9・d12・d13)に分けて表にまとめた。教科書会社ごとに述べ教材数や各教材の種類の数に偏りがないかを調べるために、全教材数と各教材の種類の数を表 2 にまとめた。表 2 より､東京書籍の全教材数のうちの評論文の割合は 0.48、小説の割合は 0.41、随想の割合は 0.11、第一学習社は評論文の割合が 0.54 小説の割合が 0.37 随想の割合が 0.09 であり、2 社に採用される教材の種類に大きな差はない。

表 2. 教材種別

	東京書籍（t1･t2・t3・t8・t10・t11）	第一学習社（d4・d5・d6・d7・d9・d12・d13）
評論文数	56	84
小説文数	48	57
随想数	13	14
全教材数	117	155

初読方法の種類を教科書ごとにまとめたところ、東京書籍で出現した初読の種類は、通読、音読、範読→通読、範読→音読、範読・音読、範読・朗読の 6 種類であった。第一学習社では、音読、通読なし、通読、音読･範読･朗読 CD、黙読、音読→黙読、黙読→音読、範読･朗読 CD、範読･音読、範読の 10 種類となった。東京書籍では｢通読｣とだけ記載することが最も多く、具体的な通読方法を指定しないことが多い。具体的な通読方法が指定されている場合もその選択理由については明記されていない。一方で、第一学習社では「音読」の指定が一番多く、通読方法の選択理由も明記されている教材が多かった。さらに「通読なし」と「黙読」の教材数が東京書籍でどちらも 0 に対し、第一学習社は前者が 10 教材、後者が 4 教材と 2 社間で差が大きかった。

初読方法の種類数が教科書会社ごとに異なることが分かったため、本研究で主に調査対象とする音読、黙読、範読、朗読 CD に結果をしぼり、音読が初読方法に含まれる教材数、黙読が初読方法に含まれる教材数、範読が初読方法に含まれる教材数、朗読 CD が初読方法に含まれる教材数をまとめた。加えて初読方法を指定しない「通読」が多かったことから､「通読」が初読方法に含まれる教材数もまとめた。まとめたものを表 3 に示す。

表 3. 教科書会社ごとの初読種類別教材数

	東京書籍 (t1・t2・t3・t8・t10・t11)	第一学習社 (d4・d5・d6・d7・d9・d12・d13)
音読を含む	8	130
黙読を含む	0	9
範読を含む	5	10
CD を含む	0	7
通読を含む	108	6

東京書籍は通読を含む初読方法が一番多く、黙読を含む初読方法と範読を含む初読方法の教材数は 0 であった。第一学習社は音読を含む初読方法が一番多く、通読を含む初読方法が 6 教材と一番少なかったが、0 の項目はなかった。

両者とも朗読 CD の付録は存在するが、東京書籍では朗読 CD や音声教材の使用が示される教材はなかった。一方で、第一学習社では指導書内で朗読 CD や音声教材の使用が示される教材として「改訂版新編現代文 A」（ｄ９）「改訂版現代文 B」（ｄ１２）「改訂版標準現代文 B」（ｄ１３）で「山月記」が、「改訂版新訂国語総合　現代文編」(ｄ４)「改訂版国語総合」(ｄ５)「改訂版標準国語総合」(ｄ６)「改訂版新編国語総合」（ｄ７）で「羅生門」が挙げられる。東京書籍で「山月記」や「羅生門」を扱わないわけではなく、「山月記」と「羅生門」を扱う場合はすべて「通読」のみの指示であった。

東京書籍と第一学習社の 2 つの教科書会社間で大きく異なる結果となったことから、教科書会社によって初読方法への考え方が異なると考えられる。例えば、2 社の大きな違いとして初読方法の選択理由の記載の有無が挙げられる。東京書籍はほとんど記載がないが、第一学習社では記載のある教材が多く見られた。学習方法を明示的にする方針であり、その際には音読が優先されるものと考えられる。

第一学習社に音読の指定が多い理由については、通読している際の指導上の留意点や通読後の活動に「語句や漢字を確認する」との指示があり、音読時の漢字等の誤読を確認する目的で設定されていることが推測される。一方、東京書籍は初読時の指導上の留意点について「構成に注意して読む」「表現・展開に注意して読む」「概要の把握」と書かれていることが多く、「語句の確認」は数教材であるが「漢字の確認」は一つもなかった。これらから、2 社間で初読時の留意点の観点が相違していることが分かる。東京書籍は初読を通して論の構成や展開といった話の流れをつかむことを目的とすることが多いのに対し、第一学習社は漢字や語句の確認といった知識の確認を目的とすることが多いと考えられる。

初読方法の理由を比較すると、音読は漢字や語句の確認のための音読、内容理解のための音読、構成の理解のための音読、さらに第一学習社の「標準現代文 B」の「経験の教えについて」（評論文）などでは「できるだけ多くの生徒に読ませ、緊張感を与える」と授業の進行に関する理由があり、緊張感を持たせるため等授業態度のための音読、文体に慣れるための音読に大きく分けられることが分かった。

黙読では理由が示されることは少ないが、漢字や語句の確認のための黙読、構成の確認のための黙読、内容理解のための黙読に分けられることが分かった。

音読、黙読の理由にある「内容理解のために」は、先行研究の高井[5]の調査において聴覚刺激により物語全体の把握が促進された影響があるという考察と近いが、高橋[1]の成人または読解力の高い読み手には音読と黙読で差がないという考察とは異なる。ここから「内容理解しやすい初読方法」については指導書の書き手の感覚によって初読方法が選択されていると考えられる。

今回の指導書の調査では教科書会社によって初読・通読方法が異なり、初読の目的も会社ごとに異なるが同一指導書内で完全に統一されているわけではないこと、全体の傾向として小説では文体や会話文に慣れるために音読が初読の方法として選択されやすく、黙読が選択されない傾向があること、評論文では黙読が選択されることが少し多いがその理由はわからないことが分かった。また、音読と朗読 CD は主に漢字や語句の確認を行う目的で選択される傾向がある。しかし初読方法の選択理由は明記されないことが多く、教科書会社が何を基準に初読方法を選択判断しているのかは断言できない。各教員が参考にする教科書会社の影響が授業に反映される可能性があるといえる。

4. 教員調査

教育現場の国語科教師がどのように初読の方法について考えているのかを調べることを目的として、アンケート調査を実施した。高等学校と中学校の国語科の教員 13 名を対象に、Microsoft Forms を利用して回答を集計した。アンケート内容を以下に示す。

1) 評論文・説明文の授業において音読・黙読・範読・朗読 CD それぞれで初読を行ったことはあるか、それぞれの選択理由は何か。
2) 物語文・小説の授業において音読・黙読・範読・朗読 CD それぞれで初読を行ったことはあるか、それぞれの選択理由は何か。

ここでは紙幅の都合上、1)、2）の初読選択理由について取りあげる。図 1 に結果を示す。図 1 の a)、b)間で分布の形状はほぼ同じであった。3.の調査から想定される初読方法の理由を選択する人が多かった。初読方法の選択理由が「特になし」とした人はどの教材でも 0 であった。

概要理解を初読方法の選択理由に挙げる人が最も多く、次に漢字語句の確認、時間の短さが続く。この順は評論文、小説ともに同じであった。また、評論文でも小説でも時間の短さを理由として音読を選んだ人はいなかった。初読方法を選択する理由に時間の短さを選ぶ人数が評論文で多い点以外、評論文と小説で大きな違いはなかった。

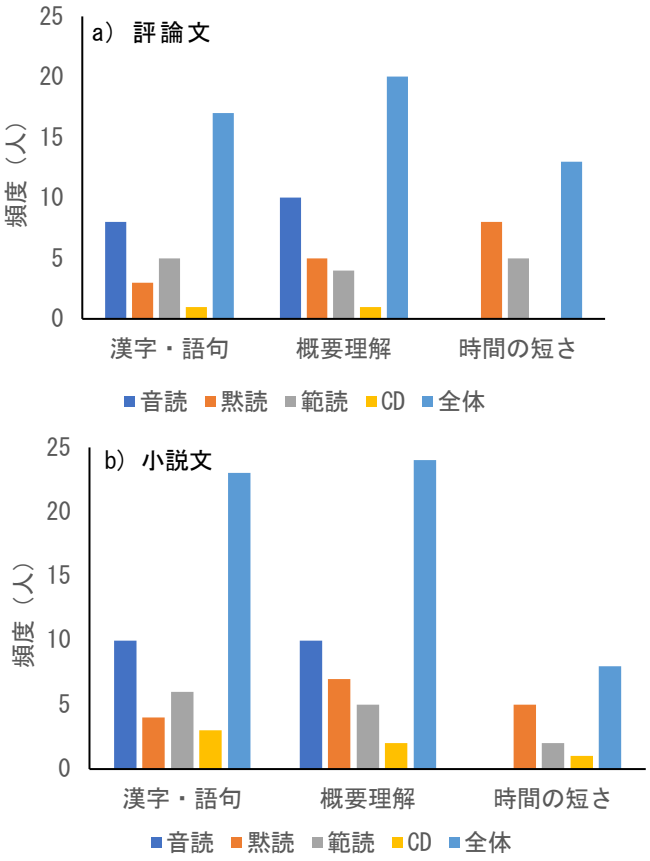


図 1. 初読方法の選択理由

5. 2 種の調査の相関性

指導書の調査と教員アンケート調査を比較して検討する。指導書の調査により、評論文では主に音読が選択されるが、小説文よりも黙読が選択されやすくなる傾向があり、一方、小説文では主に音読が選択されやすく、黙読が少ないという傾向、朗読 CD は小説に多いという結果が得られた。本稿では取り上げられなかったが、教員アンケート調査においても共通の結果が見られた。

第一学習社の指導書では初読の段階で漢字や語句の確認を目的とすることが多い結果であり、漢字や語句の確認を目的とする初読方法は、指導書の調査でも教員アンケート調査でも「音読」が選択されることが多かった。また、読む際に誤読される語句がある場合の文章に対する手立てとして「範読」が挙げられていたことも共通する。そのため、漢字や語句の確認を目的とする場合には、誤読の程度によって初読方法を変えていると考えられ、基本は音読で対応するが、文章における難解な漢字や語句が多い場合は範読にするという選択基準になっているといえるだろう。さらに、指導書の調査では、本文の漢字や語句の読みが易しいため黙読を行う教材もあったため、文章における漢字や語句が平易な場合は黙読で対応するという選択基準も考えられる。

教員のアンケート調査では内容理解を理由に初読方法を選択する人が多く、中でも音読が選ばれる傾向があったが、先行研究[6]では音読と黙読で内容理解に差がないとする考察があり、これとは異なる結果といえる。今後、詳細に検討したい。

また、教員アンケート調査においては、小説文の音読の理由に「リズムや抑揚の共有」を記述している回答者があった。これは、第一学習社の指導書において、小説文の文体や会話文に慣れることが音読の理由の一つにあることが分かった結果と同様であると考えた。また、他に、音読の理由として「寝かせない」「授業進度の統一・集中」も授業の緊張感のために選ばれている点で指導書と同じである。ここから、音読が選ばれる理由として、文体に慣れることと授業の緊張感の維持が指導書とアンケート調査に共通して得られたといえる。

今回の指導書の調査やアンケート調査から、現場では教材文の種類に限らず、音読が選択されることが多いことが分かった。また、小説における初読方法の選択理由が明示的に記述されることが評論文に比べて多く、評論文に特有の理由が明記されることが少ない。評論文で「時間の短さ」を初読法の選択理由に選ぶことが多いのも、評論文では初読が重視されていないためであると考えられる。

謝辞　東京学芸大学教育学部国語科に提出した令和 4 年度卒業論文の成果の一部である。調査にご協力くださった皆様に深謝いたします。

文　献

[1] 高橋麻衣子，人はなぜ音読をするのか―読み能力の発達における音読の役割―，教育心理学研究 61-1，95-111，2013 年.

[2] 足立幸子，初読の過程をふまえた読書指導― ハーベイ・ダニエルズ「リテラチャー・サークル」の手法を用いて，新潟大学教育学部研究紀要人文・社会科学編 6-1，1-16，2013 年.
[3] 日本国語大辞典，第 2 版，小学館，2002 年.
[4] 出版労連教科書対策委員会，教科書レポート 62，2019 年.
[5] 高井かづみ，物語の記憶・理解における呈示モダリティおよびテキストの効果，日本心理学研究 37 386-891,1989 年.
[6] 森敏昭，文章記憶に及ぼす黙読と音読の効果」教育心理学研究 28 ，57-61，1980 年.

【付録】分析対象の指導書一覧

第一学習社（2017）『高等学校改訂版新編国語総合指導と研究』
第一学習社（2017）『高等学校改訂版国語総合指導と研究』
第一学習社（2017）『第一学習社版教科書準拠高等学校改訂版国語総合朗読 CD』
第一学習社（2017）『改訂版標準国語総合指導と研究』
第一学習社（2017）『高等学校改訂版新編現代文Ａ指導と研究』
第一学習社（2018）『高等学校改訂版現代文Ｂ指導と研究』
第一学習社（2018）『高等学校改訂版標準現代文Ｂ指導と研究』
第一学習社（2018）『高等学校改訂版新訂国語総合現代文編指導と研究』
教育出版（2015）『広がる言葉小学国語 1 教師用指導書音声・動画編解説書』
東京書籍（2017）『国語総合現代文編指導書』
東京書籍（2017）『新編国語総合指導書』
東京書籍（2017）『精選国語総合指導書』
東京書籍（2018）『新編現代文 B 指導書』
東京書籍（2018）『精選現代文 B 指導書』
東京書籍（2018）『現代文Ａ指導書』

共催

日本音響学会　音声研究会

音声コミュニケーション研究会資料　(音声コミュ研資)
Proc. Tech. Comm. Speech Comm.
The Acoustical Society of Japan

発行所：　一般社団法人　日本音響学会
〒101-0021 東京都千代田区外神田 2-18-20 ナカウラ第 5 ビル 2 階
電話 (03) 5256-1020 (代)　FAX (03) 5256-1022
https://acoustics.jp/
発行人：　音声コミュニケーション研究委員会 委員長　荒井　隆行
https://asj-sccom.acoustics.jp/

発行日：　2023 年 2 月 24 日