「キミは私の声、好きかな?」大規模主観評価による声質好感度コーパスの構築とその分析

須田 仁志[1,a)]　渡邊 亞椰[2,†2,b)]　高道 慎之介[2,†1,c)]

概要：本研究では，多様な声質の音声に対して好感度の評点を与えたコーパス「CocoNut-Humoresque」を，大規模な主観評価実験により構築した．音声アナウンスや対話システムなどにおいて合成音声を利用する際には，ターゲットとなる聴取者にとって好ましい音声のデザインが有効である．本研究では，885人の聴取者に各30音声を聴取させ，総じて1800音声に対して声質にもとづく好感度の評点を収集した．話者だけでなく聴取者による好感度への影響を評価するため，聴取者に性別，年齢，好みのYouTube動画について回答させ，コーパスとして整備した．したがって本コーパスは，話者および聴取者の両側面による影響を考慮した，声質好感度の分析や推定システムの実現に貢献する．本稿では，コーパスの構築手法およびこれを用いた分析について述べ，話者および聴取者の性別や年齢に応じた好感度の傾向，また基本周波数や話者表現であるx-vectorと好感度との関係を明らかにする．

1. はじめに

音声合成技術の発展により，SiriやAlexaなどのボイスアシスタント，電話の自動応対，公共の場での自動放送，テレビやラジオでの報道など，様々な目的や場面で合成音声は利用されている [1], [2], [3], [4]．特に，テキスト音声合成（text-to-speech; TTS）や声質変換（voice conversion; VC）については，話者を表現するベクトル（話者表現）や参照音声などを入力することで，単一のモデルで多様な声質での合成を実現する手法が提案されている [5], [6], [7], [8]．これらの音声合成システムを利用すれば，目的や場面に応じた適切な声質を選択して音声を合成できる．たとえば，対話システムにおける対話相手の声質を変化させることができれば，利用者がより心地よく感じられる対話が可能になり，ユーザ体験（user experience; UX）の向上が実現できる [9], [10]．商品の広告や宣伝などの場面では，ターゲットとなる客層にとって魅力的な声質を選択することで，広告宣伝効果を増強できる [11]．あるいは，政治における選挙の候補者を宣伝する際にも，声質を適切に選択することで，政治家の信頼や説得力を向上できることが示されている [12]．しかし，音声合成システムを利用する上で音声デザインを行う際には，声質選択は設計者の主観に依存し，客観的な評価の実現は困難である．声の好感度や魅力に関する大規模なデータがあれば，統計的な分析や機械学習によるモデル化によって，より「客観的に」「主観的な」好感度を推測できることが期待される．すなわち，「その声は多くの人にとって魅力的に感じられるだろうか?」あるいは「どのような人にとってその声は魅力的に感じられるだろうか?」といった問題に対して，客観的な推測が可能になる．さらに，もし誰かに魅力的に感じてもらいたいとすれば，どのように発声すればよいかさえ推測できる．これらの統計分析や機械学習モデルは，きっとこのような質問にも答えられるであろう．——**「キミは私の声、好きかな?」「どんな声だったら、キミは好きになってくれるかな?」**

声質の魅力や好感度に関する分析については，既に多くの研究がなされている [13]．特に基本周波数 f_o と好感度の関係については複数の研究で調査されており，多人数の音声を対象とした評価も行われている．性的な魅力を対象とした評価においては，低い f_o の男性は好感度が高いことが指摘されている [14], [15]．また，低い f_o の男性は同性にとってもリーダシップが感じられることから，社会活動においても優位であることが指摘されている [16]．女性の

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Annex, AIST Tokyo Waterfront, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[†1] 現在，慶應義塾大学
Presently with Keio University
[†2] 現在，日本電信電話株式会社
Presently with NTT Corporation
[a)] suda.h@aist.go.jp
[b)] aya.watanabe@alumni.u-tokyo.ac.jp
[c)] shinnosuke_takamichi@keio.jp

声についても，適切な範囲であれば高い f_o の声が比較的好感度が高いことが指摘されている [17], [18]．これらの研究が示すように，好感度に対する f_o の影響は大きく，たとえば好みの異性を相手とした対話では，人は f_o を無意識的に変化させて発話しているとさえも指摘されている [19]．さらに，発話の音素継続長や発話速度などについても好感度に影響することが指摘されているほか，より広範な種々の音響特徴量による好感度への影響も大規模に調査されている [20], [21]．これらの研究においては，言語的に無意味なごく短時間の母音を評価させたり [14], [17], [22]，システムへの指令や同一文の発声を刺激として利用したり [18], [20] など，実用的な音声デザインとは乖離した実験条件が多く見られる．さらに，一部の研究においては合成音声や加工音声を利用した評価が行われており，自然な発話での評価が実現されていない場合がある [19], [21]．くわえて，これらの結果に反する評価結果が報告されている場合があるほか，聴取者と話者の性別の組み合わせによる影響は十分に調査されておらず，定説は確立していない [13], [22]．

本研究では，既存研究で到達していない次の 2 点に着目する．第一に，話者および聴取者の両観点において大規模な好感度評価を，自然音声を用いて行い，コーパス化する点である．発話からは，性別や年齢だけでなく，方言や話し方など様々な特徴が捉えられ，それらと好感度の関係を明らかにできれば，より緻密な音声デザインが可能になると考えられる．第二に，TTS や VC などのシステムに用いられる話者表現と好感度の関係を明らかにする点である．これらの音声合成システムでは，フォルマント周波数や声道長といったパラメータではなく，i-vector [23], d-vector [24], x-vector [25] などの連続的な話者表現が利用可能である．このような話者表現と好感度の関係を明らかにできれば，所望の好感度を持つ話者を自在に実現可能になる．特に，音声の話者の情報は，音声の音響特徴量そのもののみではなく，音響特徴量の変化にも現れることが知られている [26]．したがって，短い母音のみの音声などではなく言語的な意味を持つ発話を評価対象とすることで，その話者らしさをより確実に聴取者に評価させ，音声合成システムに入力するための話者表現との関係を明らかにすることが可能になる．

そこで本研究では，意味を成す日本語の発話に対する声質の主観的な好感度を収集し，新たなコーパスを構築した．大規模な音声コーパスを利用し，インターネット上のクラウドソーシングサービスを通じて多数の聴取者に評価させることで，多様な話者および聴取者に関する好感度の評価を実現した．さらに，音声コーパスには声質や発話スタイルなどの主観による説明文が付与されており，聴取者の感じる発話の印象が得られる．これにより，「男性の声」や「高い声」などといった単なる古典的な尺度でなく，話し方，方言やアクセント，話者の年齢，発話場面などによる好感度への影響を分析可能である．また，本コーパスには聴取者の年齢および性別の情報が含まれており，どのような聴取者が好ましく感じるかについても分析可能であり，ターゲットのユーザ属性に応じた音声デザインに貢献する．類似のコーパスとして，対話音声に対して好感度を含む種々の主観的な評価を与えた Nautilus Speaker Characterization（NSC）コーパス [27] が挙げられるが，本コーパスでは文脈に依存せずかつ意味を成す文を対象としており，対話内容についての印象を取り除いた評価を実現する．本稿では，コーパスの構築手段や構成に加え，話者および聴取者による好感度への影響を分析し，大規模なデータを活用した統計的な手段で好感度の傾向を明らかにする．本コーパスは CC BY 4.0 のもとで無償で公開されており，`https://github.com/sarulab-speech/Coco-Nut` から入手できる．

2. CocoNut-Humoresque ——大規模声質好感度コーパス——

本研究では，「誰がどのような声に対して好感度を抱くか」を詳細に調査するため，大規模な声質好感度コーパス「CocoNut-Humoresque[*1]」を構築した．1800 の各音声に 11 人以上の聴取者がそれぞれ割り当てられ，本コーパスにはそれぞれの音声–聴取者ペアについて好感度の評点が含まれている．

2.1 評価音声

評価音声として，本研究では音声表現文コーパス Coco-Nut [28] を採用した．Coco-Nut は，声に関するコメントが複数投稿された YouTube 上の動画の音声サンプルからなるコーパスである．したがって，このコーパスには，多様な声質を持つ音声が含まれており，かつ各音声が特筆されるべき特徴を持っている．Coco-Nut は 7330 の日本語の音声からなる．各音声は平均 4.0 秒であり，44 100 Hz のモノラル信号として収録されている．Coco-Nut は train, valid, test の 3 サブセットに分割されており，それぞれ 6193，559，578 の音声が含まれている．

Coco-Nut では，各音声に対して音声に関する説明文が付与されている．この説明文は，インターネット上のクラウドソーシングサービスを通じて募集した聴取者に，音声に対して「どのような話者がどのように話しているか」を主観的に記述させたものである．たとえば「40 代の男性が、優しげな声で、読み聞かせるように喋っている。」（train 0001）といった要領である．各音声に対して最低でも，train セットについては 1 文，valid セットおよび test セットについては 5 文の説明文が付与されている．なお，

[*1] 本コーパスは，Coco-Nut に含まれる音声それぞれに対して，聴取者が「すき」「きらい」をきまぐれに付与するため，CocoNut-Humoresque の名前を与えた．

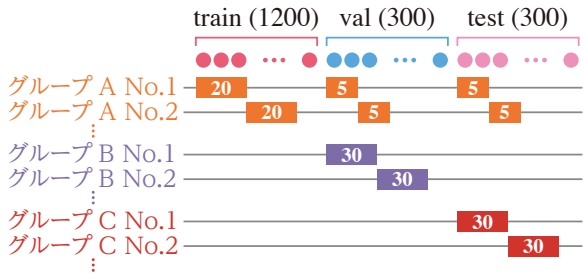


図 1 CocoNut-Humoresque における評価セット分割の模式図．各行は評価セットを，各列は音声サンプルを表す．グループ A は 60 セット，グループ B および C はそれぞれ 10 セットからなる．

Coco-Nut においては，声に対する主観的な好ましさの情報は意図的に排除されている．

Coco-Nut は，声質に関するコメントが複数投稿されるような動画から構築されている．したがって，実世界のあらゆる声質を表現したコーパスではない．一方，音声デザインなどの応用を考えれば，声質に魅力のある音声のみの収集であっても有効である．

2.2 構築手順

事前に，Coco-Nut の全音声から 80 の評価セットを作成した．これらの評価セットは，グループ A（60 セット），グループ B（10 セット），グループ C（10 セット）の 3 グループに分割されており，各セットはそれぞれ 30 音声からなる．グループ A の各評価セットには，Coco-Nut の train セット中の 20 音声，valid セットおよび test セット中のそれぞれ 5 音声が含まれている．グループ A 内の全セットを通じて，各音声は 1 セットにのみ含まれるため，総じて train セットから 1200 音声，valid セットおよび test セットからそれぞれ 300 音声が選ばれた．また，グループ A に含まれる valid セットの 300 音声を 10 セットに再分割した評価セット群をグループ B として作成した．同様に test セット中の音声に対しても処理し，グループ C の 10 セットを作成した．図 1 に，これらの評価セット分割の模式図を示す．このグループ分割は，声質の好感度を予測するシステムの評価に効果的である．たとえば，グループ A に割り当てられた聴取者については，train セット内の 20 音声に関する好感度の事前知識が与えられた条件で，test セット内の 5 音声に関する好感度の予測精度を評価できる．これに対し，グループ C に割り当てられた聴取者については，性別，年齢などのユーザ属性にもとづいた好感度の予測精度を評価できる．

評価セットは，各セット内で話者表現の観点で多様な音声が含まれるように作成した [29]．ここでは，各セットを作成する際，既にセット内に含まれるすべての音声の話者表現からユークリッド距離が最も遠い話者表現を持つ音声を選択した．このアルゴリズムを Algorithm 1 に示す．話者表現として x-vector を用い，x-vector 抽出器として WavLM を利用したモデル `microsoft/wavlm-base-plus-sv`[*2] を利用した [25], [30]．事前に x-vector の ℓ_2 ノルムが 1 となるよう正規化した．最終的に，Coco-Nut 内の説明文にもとづけば，1151 音声が男性による発話，570 音声が女性による発話であり，79 音声は説明文から推測できなかった．なお，この性別はあくまで聴取にもとづく主観的な性別であり，話者の真の性別とは異なる．したがって，本研究では，性別のバランスを保持するのではなく，x-vector の多様性の観点から音声を選択した．

Algorithm 1 サブセット S を，全音声 D から作成するアルゴリズム．

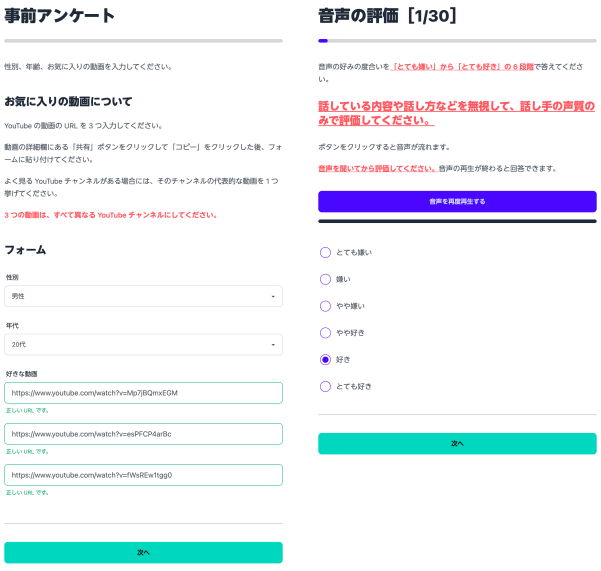
Require: サブセット内の音声数 n，全音声の集合 D，話者表現抽出器 v.
$S \leftarrow \emptyset$
$x \sim U(D)$ ▷ x は D から無作為に選択する．
while $|S| + 1 \leq n$ **do**
  $S \leftarrow S \cup \{x\}$
  $x \leftarrow \mathrm{argmax}_{x \in D \setminus S} \sum_{y \in S} \|v(x) - v(y)\|^2$
end while
return S

評価においては，まず各聴取者に性別，年齢，好みの YouTube の動画について質問した．性別については男性，女性，回答なしの 3 つの選択肢から，年齢については 10 代から 50 代までの年代と，60 歳以上，および回答なしの 7 つの選択肢から選ばせた．また，好みの YouTube の動画については，各聴取者に，YouTube チャンネルの異なる 3 つの動画の URL を入力させた．好みの動画は，聴取者の持つ文化的な背景や好ましく感じる芸能人などを収集し，それらと声質の好みの関係を調査するために収集した．

次に，各聴取者に，各音声の声質について，とても嫌い，嫌い，やや嫌い，やや好き，好き，とても好きの 6 段階で評価させた．これまでの声質の好感度評価においては，5 段階や 7 段階などの奇数段階での評価を行う例が多く見られる [21], [31]．しかし，奇数段階での評価では中間となる評点に集中し十分な好感度の情報が得られない可能性があるため，本研究では 6 段階での評価を採用した．評価の際には「話している内容や話し方などを無視して，話し手の声質のみで評価してください．」と強調して表記し，声質のみで評価するように指示した．

評価フォームは Web インタフェースとして実装した．このスクリーンショットを図 2 に示す．音声の聴取は何度でも可能としたが，それぞれの音声サンプルは必ず始めから終わりまで聴き通させた．各評価セットが均等に評価されるよう，自動で評価セットを選択し提示した．聴取者は，インターネット上のクラウドソーシングサービスであるランサーズ上で募った．各聴取者ごとに 121 円を支払い，同

[*2] `https://huggingface.co/microsoft/wavlm-base-plus-sv`



(a) 事前アンケートフォーム　　(b) 評価フォーム

図 2 評価のための Web フォームのスクリーンショット．(a) の画面で事前アンケートに回答したのち，(b) の画面で音声を評価する作業を 30 回繰り返す．

表 1 性別・年代ごとの聴取者の数．N/A は無回答を表す．

	10s	20s	30s	40s	50s	60–	N/A	合計
男性	7	58	137	189	93	41	1	526
女性	1	45	97	131	64	15	2	355
N/A	0	1	1	0	0	0	2	4
合計	8	104	235	320	157	56	5	885

一の聴取者による複数回の参加は禁止した．

2.3 聴取者の属性

評価に参加した聴取者は 885 人である．80 サブセットそれぞれを均等に評価させたため，各サブセットは少なくとも 11 人の聴取者により評価された．すべてのサブセットについて，少なくとも 2 人の男性および女性の聴取者が割り当てられた．表 1 に性別・年代ごとの聴取者の数を示す．およそ 59%の聴取者が男性であり，40%の聴取者が女性であった．また，平均年齢はおよそ 40 歳であった．総じて，10 代の参加者はごく少ないものの，多様な性別・年代の聴取者が評価に参加した．なお，これらの聴取者の性別・年代の分布は，クラウドソーシングによる評価のバイアスの存在を示唆している．したがって，クラウドソーシングを通じたあらゆる評価においては，評価が性別や年齢に依存する可能性およびその分布が実世界と異なる可能性を考慮すべきである．

聴取者の好む YouTube 動画について，動画に付与されているカテゴリ別の動画数を図 3 に示す[*3]．最も多いカテゴリは「エンターテイメント」であり，様々な嗜好を持つ聴取者が参加したことが示唆された．また，これらの YouTube 動画に付与されているタグについて，ワードクラウドとして図 4 に可視化した[*4]．バーチャル YouTuber，ゲーム実況，音楽，学びや投資など，様々な属性の YouTube 動画が含まれていることが示された．

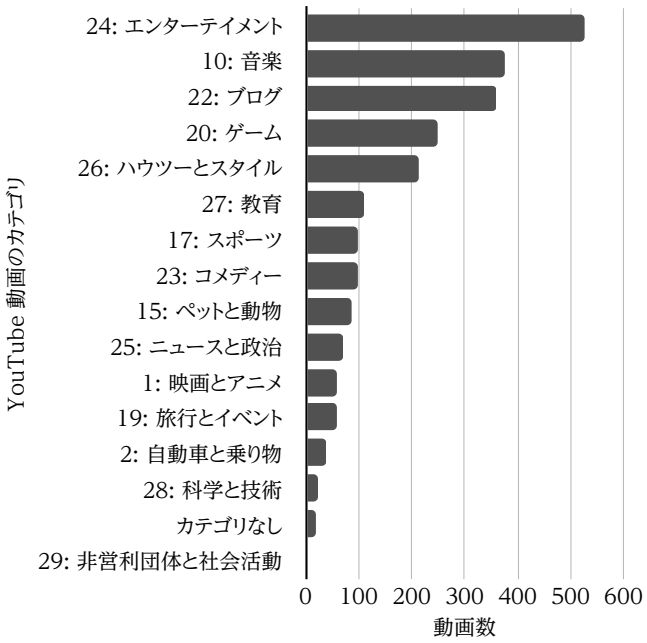


図 3 聴取者の好みの YouTube 動画の，カテゴリ別の合計数．一部の聴取者は不適切なフォーマットで回答したため，合計は規定の数（2655）より少ない．

図 4 聴取者の好みの YouTube 動画に付与されているタグから生成したワードクラウド．単語の大きさは頻度を示す．

[*3] YouTube におけるカテゴリは，各動画に 1 つのみ設定可能なメタデータである．およそ 30 カテゴリが存在するが，一般の投稿者はこのうち 15 カテゴリから選択できる．なお，カテゴリの情報は，YouTube のインタフェース上で直接確認することはできないようである．

[*4] YouTube におけるタグは，投稿者が各動画に複数設定可能なメタデータである．

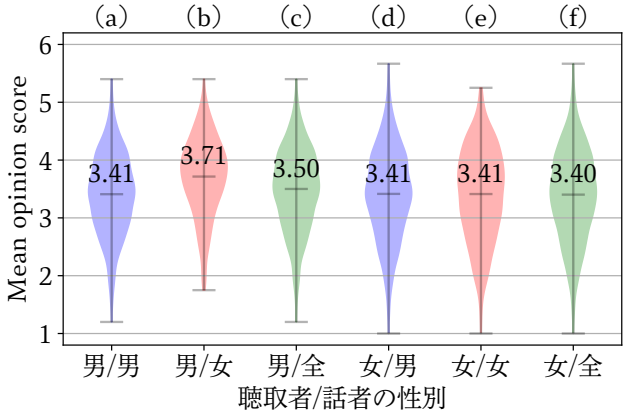


図 5 聴取者・話者ごとの MOS の分布．各列のラベルの左が聴取者の性別を，右が話者の性別を表す．「全」は全話者についての分布を示す．このうち，(a)–(b)，(a)–(c)，(b)–(c)，(b)–(d)，(b)–(e)，(b)–(f)，(c)–(d)，(c)–(e)，(c)–(f) に $p < 0.05$ で有意差が見られた．

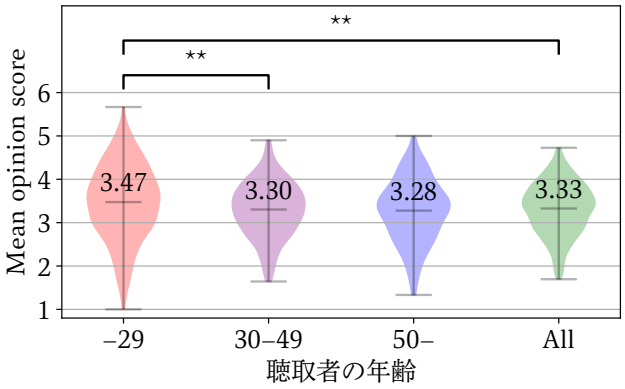


図 6 聴取者の年代ごとの MOS の分布．図中の**は，$p < 0.01$ で有意差が見られた組を示す．

3. 分析 1: 性別・年齢による好感度への影響

話者および聴取者の性別・年齢ごとの好感度の傾向を調査するため，各カテゴリ・発話ごとに平均オピニオン評点（mean opinion score; MOS）を計算し，その分布を調査した．

3.1 性別による影響

話者および聴取者の性別による評点への影響を分析した．話者については，Coco-Nut 内に含まれる表現文に「男」もしくは「女」が含まれる場合，その性別を話者の性別と見做した．性別に関する情報が含まれていない場合，また複数の表現文に異なる性別が含まれる場合には，話者の性別を不明として扱った．2.2 節で記述したとおり，この性別は聴感上の性別であり，話者の真の性別とは異なる可能性がある．

図 5 に MOS の分布のバイオリンプロットを示す．話者の性別を無視すれば，図 5 の (c) と (f) を比較すると，女性の聴取者が与える評点は，低くかつ分散している．それぞれ，Welch の t 検定にもとづけば p 値は 7.0×10^{-5} 未満，Bartlett 検定にもとづけば p 値は 1.9×10^{-5} 未満で有意な差が認められた．さらに，聴取者が男性の場合，図 5 の (a) と (b) を比較すると，Welch の t 検定にもとづいて $p < 1.0 \times 10^{-5}$ で，女性話者の発話への評点が有意に高い．一方，聴取者が女性の場合，図 5 の (d) と (e) を比較すれば，話者の性別による有意な評点の差は見られない．これらの結果より，聴取者・話者による評点の差は，男性の聴取者が女性話者の声を評価する場合に有意に高くなる，という場合を除き大きな差は生じないことが結論づけられる．

既存研究では，好感度の傾向は，各性別内では多く議論されているものの，異なる性別間での差は明らかにされていない [13]．本評価で性差が現れた要因として，Coco-Nut に含まれる女性の声が総じて男性に魅力的であった可能性や，好感度そのものではなく評点を与える傾向の性差などが考えられ，あらゆる女性の声に対して男性が高い好感度を持つとはただちに結論付けられない．ただし，本節の議論から評点上の明らかな性差が生じることが確認されたため，音声デザインの評価においてこの性差について考慮する必要があることは確かである．特に，既存のボイスアシスタントなどでは，ジェンダーバイアスの観点での問題は無視できないものの，女性の話者が既定で選ばれている場合が多い [32]．本節で議論したような評点上の性差の観点からも，女性話者を選択することは製品の数値上の評価を向上させるためには——あくまで数値的な評価ではあるが——合理的であると考えられる．

3.2 聴取者の年齢による影響

聴取者の年齢による評点への影響を分析した．ここでは，データ数の不足を考慮し，30 歳未満，30 代もしくは 40 代，および 50 歳以上の 3 グループに聴取者を区分けした．また，評点のサンプルが多い valid セットおよび test セットの音声のみを分析対象とした．本コーパスの設計上，それぞれの発話は 22 人以上の聴取者によって評価されている．なお，いずれかの性別・年代グループに属する聴取者がいない場合，その発話を無視した．図 6 に MOS の分布のバイオリンプロットを示す．30 歳未満の聴取者は，30 歳以上の聴取者と比較して，Welch の t 検定にもとづけば p 値は 0.01 未満で，有意に高い評点を与えた．分散についても，Bartlett 検定にもとづけば p 値は 10^{-21} 未満で，有意に大きいことが確認された．

くわえて，聴取者の性別・年齢の組み合わせによる評点への影響を調査した．図 7 に MOS の分布のバイオリンプロットを示す．女性聴取者では有意差が見られないものの，前述と同様に，30 歳未満の聴取者が，平均が高く分散が大きい評点を与える傾向があることが示された．この傾

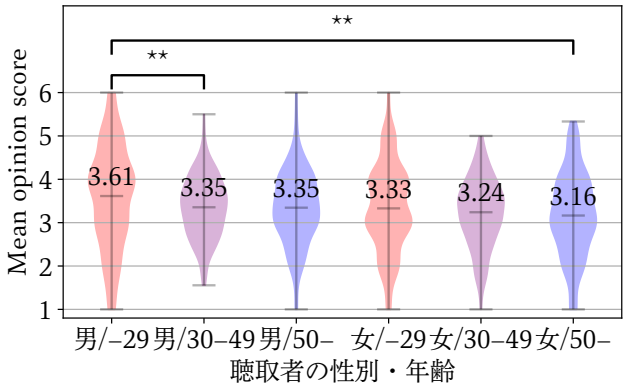


図 7 聴取者の性別・年代ごとの MOS の分布．各列のラベルの左が聴取者の性別を，右が年代を表す．図中の**は，$p < 0.01$ で有意差が見られた組を示す．

向は，話者が男性であっても女性であっても共通して見られる傾向であった．

3.3 本節の総括

本節の議論の結果は，以下の 3 点に要約される．

(1) 年齢を無視した際，男性聴取者が女性話者に与える評点は高く，それ以外の組み合わせの間では大きな差が見られない．

(2) (1) の状況から，平均すれば男性聴取者は高い評点を与えやすい．

(3) 30 歳未満の男性の聴取者による評点は，話者の性別に関係なく，平均が高く分散が大きい．

なお，これらの要点にもとづけば，30 歳未満の男性が与えた評点は，50 歳以上の女性が与えた評点を大きく上回ると推測される．事実，図 7 に示すように，これらの平均評点には 0.45 もの差があり，Welch の t 検定にもとづけば $p < 10^{-5}$ で有意な差が見られた．

4. 分析 2: サンプルごとの好感度分析

4.1 特定の性別のみが好感度を抱く声質

サンプルごとの分析により，男性もしくは女性の聴取者のみが特に好感度を抱く音声を明らかにできる．この考察は，合成音声の対象聴取者と音声デザイナーとで性別が異なる場合の，主観的な認知のへだたりの抑制などに活用できる．この分析においては，評点のサンプルが多い valid セットおよび test セットの音声のみを対象とした．

図 8 に，MOS の分布と聴取者の性別との関係を示す．男女間の評点の相関係数は 0.65（$p < 10^{-73}$）であり，男女の評点に大きな相関が見られた．一方，男女間で大きく評価が分かれた音声も見られた．図 8 (a) は，Coco-Nut 中の valid セットに含まれる第 7230 文で，「若い男性が、はきはきした低い声で、怒ったように喋っている。」「若い男性が芝居がかった声で早口で喋っている。」などと表現

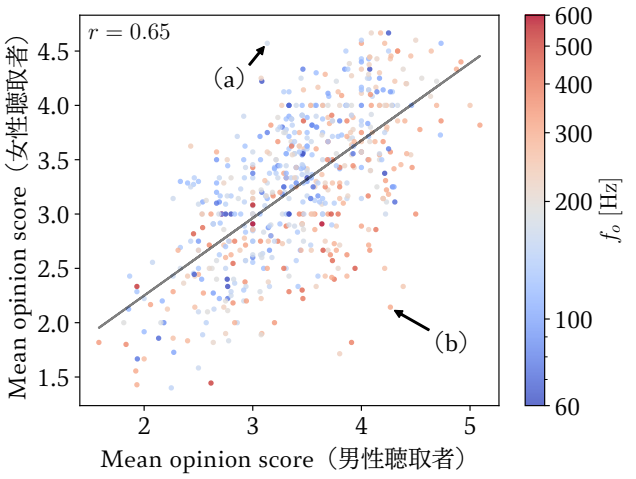


図 8 男性聴取者と女性聴取者による MOS にもとづく散布図．図中の左上が女性によく好まれるサンプルを，右下が男性によく好まれるサンプルを表す．図中の (a) および (b) は男女で最も MOS の差が大きいサンプルを示す．基本周波数 f_o は，Crepe の full model [33] により得られた対数基本周波数系列の平均から得た．

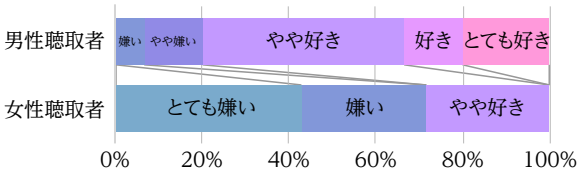


図 9 図 8 (b) で示した test セット第 6979 文についての男女別の好感度の評点．平均オピニオン評点は，男性聴取者は 4.27 で女性聴取者は 2.21 である．この音声については，男性聴取者 15 人，女性聴取者 7 人により評価された．

されており，本評価の中でとりわけ女性に好まれた[*5]．一方，図 8 (b) は，Coco-Nut 中の test セットに含まれる第 6979 文で，「10 代後半から 20 代前半くらいの女性が、落ち着いた幼気な声で、ぼそぼそとつぶやくように喋っている。」「10 代の少女が、かわいらしい声で、まったりとした口調で喋っている。」などと表現されており，本評価の中でとりわけ男性に好まれた[*6]．特に図 8 (b) の音声については，図 9 に示すように，男性聴取者の 80%が「やや好き」以上と答えた一方，女性聴取者の 71%が「やや嫌い」以下と答えた．したがって，聴取者の性別のみにおいてもサンプルごとに好感度の傾向の違いが見られ，特定の性別のみにとりわけ好まれる声質の選択が可能であることが示唆された．それと同時に，対象聴取者と音声デザイナーとの性別が異なる音声デザインにおいては，音声デザイン時に考慮できないほどの好感度の大きな性差が現れる場合がある点について認識する必要がある．

[*5] 著者の聴感上は，20 代のアニメキャラクターのような男性が大声で驚いているような音声である．

[*6] 著者の聴感上は，20 代前半のゲームキャラクターのような女の子が，聞き手を優しく受け入れるような音声である．

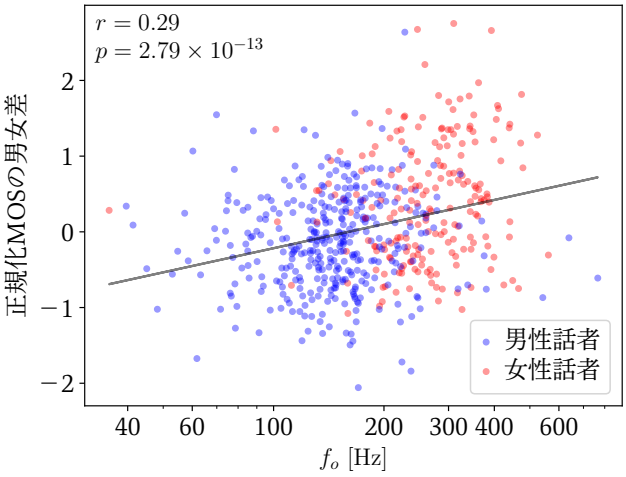


図 10 基本周波数 f_o と，男女間での好感度の評点の差の関係．縦軸は，男性聴取者の正規化 MOS から女性聴取者の正規化 MOS を減算したものである．すなわち，図中でより上のサンプルが，より男性に特異的に好まれるサンプルである．男性聴取者と女性聴取者の間で MOS の分布に違いが生じるため，図 5 の結果に従い，それぞれ平均が 0 かつ分散が 1 となるよう線形に正規化した．

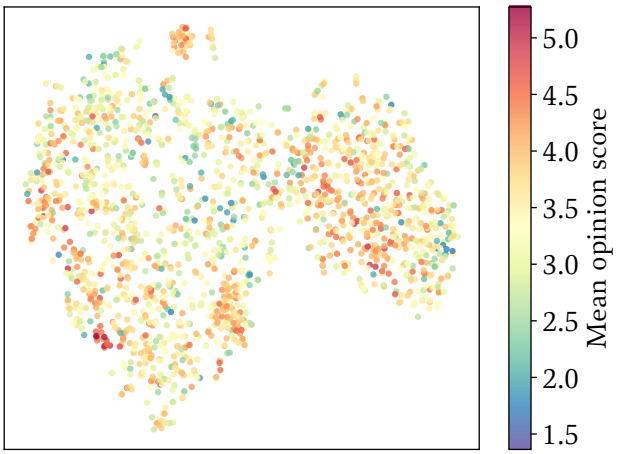
表 2 すべての聴取者・話者の性別の組み合わせにおける，対数 f_o と好感度の正規化 MOS の相関係数．*で示した相関は $p < 0.05$ で有意であり，**で示した相関は $p < 0.01$ で有意である．

		聴取者		
		男性	女性	すべて
話者	男性	−0.133*	−0.169**	−0.164**
	女性	0.078	−0.067	0.003
	すべて	0.139**	−0.104*	0.020

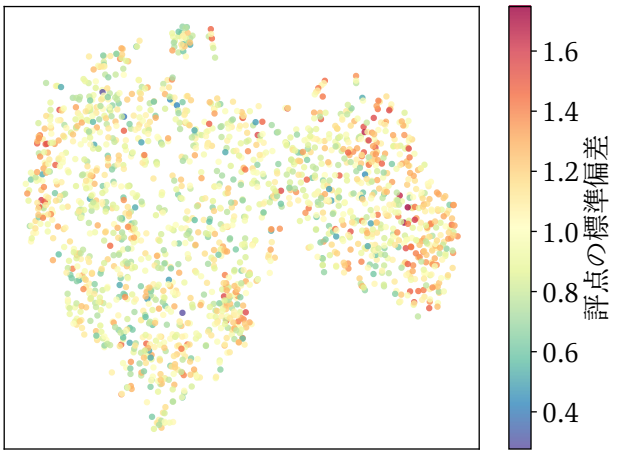
4.2 基本周波数 f_o と好感度の関係

各音声の f_o と，その音声の聴取者についての評点の性差について，図 10 に可視化する．ただし，ここでは男女の聴取者での評点の傾向の違いを抑制するため，聴取者の性別に応じて MOS を平均 0 かつ分散が 1 となるように線形に正規化した．大域的には，低い男性の発声は女性により好まれ，高い女性の発声はより男性に好まれることが，有意な相関として確認された．一方，各音声について個別に観察すれば，低い男性の音声においてもより男性に好まれる場合があるほか，その逆の場合も見られる．したがって，聴取者についての評点の性差は f_o のみでは決定できない．

聴取者・話者についてのすべての性別の組み合わせにおいて，対数 f_o と好感度の相関を調査した結果を表 2 に示す．男性話者については，聴取者の性別によらず，低い f_o の話者に高い好感度が得られることが示された．一方，女性話者については，聴取者の性別によらず大きな相関は見られなかった．また総じて，男性の聴取者は高い f_o の話者に，女性の聴取者は低い f_o の話者に好感度を持つことが示された．この結果は図 5 や図 10 の結果にも関連する．



(a) MOS



(b) 評点の標準偏差

図 11 MOS もしくは評点の標準偏差により色付けした，x-vector の t-SNE による可視化．これらの図においては，男性話者の発話が左側のクラスタに，女性話者の発話が右側のクラスタにおよそ属している．

1 節で述べたとおり，音声の f_o と好感度の関係は既に広く研究されている．特に男性話者については，低い f_o の話者は高い好感度を得られる傾向にあることが種々の研究で示されている [14], [15], [16]．表 2 に示すように，この傾向は日本語についても強く確認された．

4.3 好感度と x-vector の関係

本節では，話者表現である x-vector [25] と好感度との関係について議論する．ここでは，2.2 節の場合と同様に，WavLM を利用した x-vector 抽出モデルである `microsoft/wavlm-base-plus-sv`[*7]を利用した [30]．得られた x-vector を t-SNE によって 2 次元に圧縮し可視化した結果を図 11 に示す．図 11 においては，話者の性別によりおよそのクラスタが形成されており，図中左のクラスタが男性話者，右のクラスタが女性話者の発話から主に形成されている．図 11 から，x-vector 空間内の位置により好感度が異なることが示唆された．たとえば，図中最左部分（よ

[*7] `https://huggingface.co/microsoft/wavlm-base-plus-sv`

り男性らしい男性の声）や，右のクラスタの左部分（中性的な女性の声）が比較的 MOS が高いことが観察される[*8]．特に前者については，表 2 に示した男性話者の好感度の傾向と合致する．また，図中最右部分（特に女性らしい女性の声）などは評点の標準偏差が大きく，好みが分かれる音声についても x-vector 空間内で集まる傾向にあった．

4.4 本節の総括

本節の議論の結果は，以下の 4 点に要約される．

(1) f_o の低い男性は好感度が高く，女性の好感度と f_o の関係は確認できない．
(2) f_o の高い話者は男性にとって好感度が高く，f_o の低い話者は女性にとって好感度が高い．
(3) (1) および (2) の結果はあくまで大域的であり，好感度の傾向はそれぞれの話者によって大きく異なる．
(4) x-vector 空間内での位置と好感度に関係が見られ，x-vector から好感度の傾向は予測可能であると示唆される．

したがって，f_o や x-vector などの既存の特徴量を利用することで，特定の対象の聴取者にとって好感度の高い音声をデザインできる可能性が示された．また，個別のサンプルに着目すれば，全聴取者による評点の標準偏差が非常に低い場合があり，どのような聴取者にとっても好感度の高い音声あるいは低い音声をデザインできる可能性についても示唆された．

5. おわりに

本研究では，ボイスアシスタントや商品の広告・宣伝などに用いる合成音声のデザインを支援するため，音声の主観評価による声質好感度コーパス「CocoNut-Humoresque」を構築した．これは，1800 の音声にそれぞれ 11 人以上の聴取者を割り当て評価させた，総じて 885 人の聴取者による大規模なコーパスである．本コーパスの対象音声は YouTube 上から収集された多様な声質の話者による音声であり，各音声に対して表現文が付与されているため，話者の詳細な印象と好感度との関係を見出すことができる．各聴取者には，性別，年齢のほかに好みの YouTube 動画を回答させ，ユーザの持つ属性や文化的背景などと好感度との関係を分析可能にした．本コーパスの分析により，男性聴取者は女性話者に対して高い評点を与える傾向にあるなど，聴取者–話者間の性別および年齢の組み合わせによる評点への影響を明らかにした．また，各音声それぞれについて分析を行い，男性話者のうち f_o が低い話者ほど高い好感度を得やすいなどの大域的な傾向や，好感度の性差が大きく生じる話者の存在などを明らかにした．さらに，話者表現である x-vector と好感度との関係を明らかにし，x-vector 空間内での位置から好感度の評点の平均や分散が推測可能であることを示した．

本研究では，音声からの好感度予測や，特定のユーザ属性にターゲティングした声質選択支援など，実システムの構築には至っていない．特に好感度予測については，音声の品質に対する MOS の予測手法を応用可能である [34], [35]．しかし，音声品質の評価と異なり，好感度は聴取者によって大きく異なるため，その好感度の違いを考慮したモデルを構築する必要がある．また，好感度の予測結果の評価手法についても自明でなく，個別の考察により評価基準を設定する必要がある．さらに，音声デザインを支援する声質選択システムについては，最終的には実プロダクトにおける効果の測定が求められる．声質選択の効果を定量的に評価することは単純でなく，システム構築の先の複数の事例収集を交えた考察が必要である．

謝辞 本研究は JSPS 科研費 23K20017，21H04900，22H03639，23H03418，JST 創発的研究支援事業 JPMJFR226V の助成を受けたものです．この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP20006）の結果得られたものです．

[*8] 男性らしさや女性らしさについての主観的な評価は行われておらず，あくまで x-vector 上での位置関係によってのみ推測した．

参考文献

[1] Hoy, M. B.: Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, *Medical reference services quarterly*, Vol. 37, No. 1, pp. 81–88 (2018).
[2] Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C. and Voße, J.: Speech synthesis evaluation — State-of-the-art assessment and suggestion for a novel research program, *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)* (2019).
[3] 木田祐介，藤田雄介：LINE CLOVA の音声認識技術，研究報告音楽情報科学（MUS），Vol. 2023-MUS-137, No. 2, pp. 1–1 (2023).
[4] 栗原　清：日本語音声合成を用いた AI アナウンスシステムの研究と実用化，映像情報メディア学会誌，Vol. 78, No. 2, pp. 234–242 (2024).
[5] Toda, T., Ohtani, Y. and Shikano, K.: One-to-many and many-to-one voice conversion based on eigenvoices, *Proc. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, pp. IV–1249–IV–1252 (2007).
[6] Arık, S. Ö., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J. and Zhou, Y.: Deep voice 2: multi-speaker neural text-to-speech, *Proc. 31st International Conference on Neural Information Processing Systems*, pp. 2966–2974 (2017).
[7] Chou, J.-C. and Lee, H.-Y.: One-shot voice conversion by separating speaker and content representations with instance normalization, *Proc. Interspeech 2019* (2019).
[8] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E. and Ponti, M. A.: YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, *Proc. 39th International Conference on Machine Learning*, Vol. 162, pp. 2709–2720 (2022).
[9] Yu, Q., Nguyen, T., Prakkamakul, S. and Salehi, N.: "I almost fell in love with a machine": Speaking with com-

puters affects self-disclosure, *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, No. Paper LBW0255, pp. 1–6 (2019).
[10] Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M. and Bernstein, A.: Female by default? — Exploring the effect of voice assistant gender and pitch on trait and trust attribution, *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, No. Article 455, pp. 1–7 (2021).
[11] Burkhardt, F., Huber, R. and Batliner, A.: Application of speaker classification in human machine dialog systems, *Speaker classification I: Fundamentals, features, and methods* (Müller, C., ed.), Berlin, Heidelberg, pp. 174–179 (2007).
[12] 岡田陽介：政治家の印象形成における声の高低の影響: 音声合成ソフトを用いた女声による実験研究，応用社会学研究， Vol. 58, pp. 53–66 (2016).
[13] Weiss, B., Trouvain, J., Barkat-Defradas, M. and Ohala, J. J.(eds.): *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, Springer Nature Singapore (2020).
[14] Skrinda, I., Krama, T., Kecko, S., Moore, F. R., Kaasik, A., Meija, L., Lietuvietis, V., Rantala, M. J. and Krams, I.: Body height, immunity, facial and vocal attractiveness in young men, *Die Naturwissenschaften*, Vol. 101, No. 12, pp. 1017–1025 (2014).
[15] Suire, A., Raymond, M. and Barkat-Defradas, M.: Human vocal behavior within competitive and courtship contexts and its relation to mating success, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 39, No. 6, pp. 684–691 (2018).
[16] Mayew, W. J., Parsons, C. A. and Venkatachalam, M.: Voice pitch and the labor market success of male chief executive officers, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 34, No. 4, pp. 243–248 (2013).
[17] Borkowska, B. and Pawlowski, B.: Female voice frequency in the context of dominance and attractiveness perception, *Animal behaviour*, Vol. 82, No. 1, pp. 55–59 (2011).
[18] Puts, D. A., Barndt, J. L., Welling, L. L. M., Dawood, K. and Burriss, R. P.: Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness, *Personality and individual differences*, Vol. 50, No. 1, pp. 111–115 (2011).
[19] Jones, B. C., Feinberg, D. R., Debruine, L. M., Little, A. C. and Vukovic, J.: Integrating cues of social interest and voice pitch in men's preferences for women's voices, *Biology letters*, Vol. 4, No. 2, pp. 192–194 (2008).
[20] Burkhardt, F., Schuller, B., Weiss, B. and Weninger, F.: "Would you buy a car from me?" — On the likability of telephone voices, *Proc. Interspeech 2011* (2011).
[21] Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D. and Delplanque, S.: Voice attractiveness: Influence of stimulus duration and type, *Behavior research methods*, Vol. 45, No. 2, pp. 405–413 (2013).
[22] Zheng, Y., Compton, B. J., Heyman, G. D. and Jiang, Z.: Vocal attractiveness and voluntarily pitch-shifted voices, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 41, No. 2, pp. 170–175 (2020).
[23] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).
[24] Variani, E., Lei, X., McDermott, E., Moreno, I. L. and Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification, *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056 (2014).
[25] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker recognition, *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333 (2018).
[26] Furui, S.: Cepstral analysis technique for automatic speaker verification, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 29, No. 2, pp. 254–272 (1981).
[27] Fernández Gallardo, L. and Weiss, B.: The Nautilus Speaker Characterization Corpus: Speech recordings and labels of speaker characteristics and voice descriptions, *Proc. 11th International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
[28] Watanabe, A., Takamichi, S., Saito, Y., Nakata, W., Xin, D. and Saruwatari, H.: Coco-Nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control, *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2023).
[29] Seki, K., Takamichi, S., Saeki, T. and Saruwatari, H.: Diversity-based core-set selection for text-to-speech with linguistic and acoustic features, *Proc. 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12351–12355 (2024).
[30] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X. and Wei, F.: WavLM: Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, pp. 1505–1518 (2022).
[31] Hughes, S. M., Dispenza, F. and Gallup, G. G.: Ratings of voice attractiveness predict sexual behavior and body configuration, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 25, No. 5, pp. 295–304 (2004).
[32] Loideain, N. N. and Adams, R.: From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of data protection impact assessments, *Computer Law & Security Review*, Vol. 36, p. 105366 (2020).
[33] Kim, J. W., Salamon, J., Li, P. and Bello, J. P.: Crepe: A convolutional representation for pitch estimation, *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165 (2018).
[34] Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R. A. and Sculley, D.: AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech, *Proc. NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop* (2016).
[35] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S. and Saruwatari, H.: UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022, *Proc. Interspeech 2022*, ISCA (2022).