人と AI のあいだにある声
～ ハイブリッド音声が醸し出す「親しみ」の分析 ～

何　平[†1]　武藤　剛[†2]

概要：本研究では，特定話者の音声と特定話者が存在しない人工知能（AI）により生成された人工音声を任意の比率で混合した「ハイブリッド音声」を生成するシステムを新たに構築し，それにより生成された人工音声に対する聴取印象がどのように変化するかの検討を行った．具体的には，高精度音声合成モデル XTTS-v2 を用いて，面識のある人物（知人）の参照音声から生成した日本語と中国語の母語話者の音声と，モデルが生成する人工音声を複数の混合比率で作成し，評価刺激として提示した．そして，その音声を中国語母語話者 10 名および日本語母語話者 9 名を対象に実施し，混合比率の違いおよび聴取者と話者の親密度が印象形成に与える影響を分析した．その結果，人工音声とのバランスが印象形成に関与する可能性が示された．特に，混合比率が 0.5 において「違和感」や「よそよそしさ」は相対的に少なく，「好感」および「親近感」は相対的に高まる傾向が示され，知人の音声と人工音声との中庸なバランスが，「親しみ」のある声という印象形成に関与する可能性が示唆された．このことは，本提案手法が，対話型エージェントや感情共感型 AI において，ユーザ特性や利用文脈に応じた合成音声の設計指針となる可能性を示している．

1. はじめに

近年，音声合成技術は急速に発展しており，とりわけ多言語・多話者条件で高い再現性を示す XTTS シリーズの登場により，特定話者の音声データからその声質や韻律的特徴といった話者性を高精度に再現した合成音声である「クローン音声」の生成が可能となった．XTTS-v2 は従来の TTS より自然性と滑らかさに優れ，多様な話者・言語条件下でも安定した音質を保つ点で注目されている[1]．こうしたクローン音声技術の進展は，個人化音声エージェントや感情表現型 AI の設計に大きな可能性をもたらしている．

加えて，対話型システムにおける音声は「情報提示」だけでなく，利用継続や信頼感・親密感といった関係性の形成にも関与するため，同一内容の発話であっても，声質・韻律・話者らしさがユーザ体験に影響し得る．そのため，音声の「人間らしさ」を高めるだけでなく，目的（親近感の喚起，違和感の低減など）に応じて声の印象を設計・制御する枠組みが重要になる．

一方で，合成音声の知覚に関する感性的側面は十分に解明されておらず，とりわけ合成音声の聴取印象変化に関する体系的知見は乏しい．既存研究は主に「自然さ」，「好感度」「違和感」などの主観評価に焦点を当てており[2]，合成音声がもつ魅力や人間らしさが聴取者の信頼感に影響することも指摘されている[3]．また，合成音声の人間らしさを技術的に高めることで，自然さや好感度が一貫して向上する傾向も報告されている[4]．

しかし，単に「より自然にする」だけでは，必ずしも関係性の形成に直結しない可能性がある．たとえば，クローン音声は話者同一性が高い反面，聴取状況や提示文脈によっては過度に生々しく感じられ，違和感や距離感につながることがある．一方で，特定話者が存在しない人工知能（AI）による音声合成によって生成された音声（本研究では「人工音声」と呼ぶ）は，均質で安定した品質を得やすい一方で，個人性の希薄さが「よそよそしさ」として知覚される可能性がある．このように，話者同一性（本人らしさ）と人工性（作為感）のバランスを調整することが，印象設計の観点から有効な操作因子になり得る．

クローン音声や参照音声を用いる方式では，参照音声の録音品質（環境雑音・残響・風切り音など）が推定される話者特徴に影響し，結果として合成音声の安定性や印象評価にも影響を与え得る．単一マイク条件の雑音環境に対しては，短時間フーリエ変換（STFT）領域で雑音パワースペクトルを推定し，ゲイン関数で抑圧する統計的音声強調が古典的に用いられてきた．特に，対数スペクトル誤差に基づく MMSE log-STSA 系の推定は，音声処理に適した歪み尺度として広く参照される[5]．さらに，非定常雑音に対しては，音声存在確率を用いて雑音推定を追従させる MCRA と，音声存在不確かさを考慮した OM-LSA により，弱い音声成分を保持しつつ“musical noise”を回避する枠組みが提案されている[6]．また，IMCRA は最小値追跡を二段で行うことで，低 SNR・非定常雑音・弱い音声成分を含む条件でも雑音推定を頑健化し得ることが示されている[7]．

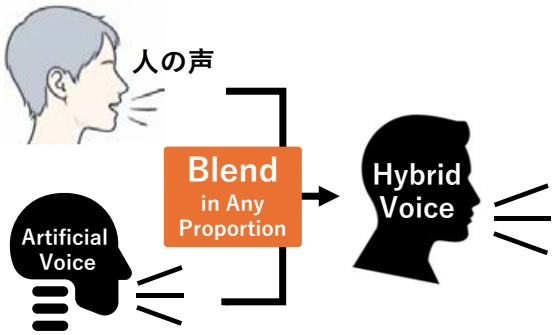

図 1 ハイブリッド音声の概要

＊†1 文教大学大学院　情報学研究科
＊†2 文教大学　情報学部

本研究のように「話者特徴の保持」が重要な用途では，過度な歪みを避けつつ参照音声の不要成分を抑える前処理が，実験刺激の品質管理として重要になる．

そこで，本研究は，特定話者の音声と人工音声を任意の比率で混合（Blend）できるハイブリッドの合成音声（以下，ハイブリッド音声）の作成手法（図 1）を新たに考案し，関係性の形成に寄与できる「親しみ」のある合成音声の生成手法の構築を目的とする．具体的には，XTTS-v2 のような高精度モデルを用いて条件を統一し，面識のある人物（知人）の音声と人工音声を様々な比率で混合したハイブリッド音声が聴取印象に与える影響を明らかにする．評価指標として「聞き取りやすさ」，「好感」，「親近感」，「違和感」，「よそよそしさ」の 5 項目を用い，混合比率および聴取者と話者の関係が，「親しみ」の印象形成に及ぼす効果を分析する．さらに，言語特性の影響を考慮するため，日本語および中国語の母語話者を対象に独立に実験を行い，言語環境を跨いだ傾向の比較検討も行う．

2. 手法

2.1 音声生成

XTTS-v2（Coqui GmbH，2023）を用い，Python（PyTorch 2.1，CUDA 環境）上に生成スクリプトを構築した．テキスト入力から音声出力までの処理を自動化し，サンプリングレートは 22.05 kHz に統一した. 音声サンプルの記録には，日本語母語話者および中国語母語話者をそれぞれ 1 名ずつ参加し，それぞれ約 10–20 秒の音声を収録して XTTS-v2 によりクローン化する．以降，特定話者の音声に基づく合成音声を「クローン音声」，XTTS-v2 で生成した基準的な人工音声を「人工音声」と呼ぶ．

参照音声（クローン化用サンプル）は録音環境雑音（例：ファンノイズや空調音など）を含み得るため，話者特徴推定の安定化および刺激品質の統一を目的として，必要に応じて降噪前処理を行った. 単一マイク条件の音声強調では，雑音パワースペクトル推定とゲイン関数に基づく抑圧が基本構成となり，log-STSA に基づく推定が古典的枠組みとして知られている[5]．また，非定常雑音下では，雑音推定を追従させる MCRA と推定器である OM-LSA を組み合わせ，弱い音声成分の保持と musical noise の抑制を両立する枠組みが提案されている[6]．さらに IMCRA は，最小値制御再帰平均の改良により，低 SNR・非定常雑音・弱い音声成分を含む条件でも雑音推定を頑健化し得る[7]．本研究ではこれら先行研究の考え方を踏まえ，「話者性の保持」を優先しつつ，顕著な環境雑音が含まれる場合に限り，過度な歪みを避ける範囲で抑圧を行う設計とした（パラメータは保守的に設定）．また，言語比較の公平性を確保するため，生成環境（同一デバイス・同一モデル・同一サンプリングレート）を固定し，出力 WAV のフォーマットも統一した．これにより，聴取印象の差が主として「混合比率」と「言語条件」に由来するよう統制した．

参照音声の降噪は，本研究では操作の直感性を優先し，以下のように「録音から推定した雑音成分を差し引く」処理として表現する．すなわち，原録音 x から推定雑音 $\hat{n}$ を強度係数 α で制御して差し引き，降噪後信号 y を式（1）に示す．

$$y = x - \alpha\hat{n} \tag{1}$$

ここで α は「降噪強度」に相当し，値が大きいほど雑音抑圧が強くなる．

一方，雑音を強く抑圧すると水声（musical noise）や声質の変化が生じ得るため，音色の再現性を確保する目的で，降噪後信号 y と原信号 x を混合係数 d で線形合成し，最終出力 out を式（2）に示す．

$$out = dx + (1-d)y \tag{2}$$

なお，混合係数 d が大きいほど原音が多く残り，歪みを抑えやすい反面，雑音残留が増える．この前処理は，主として参照音声（クローン化用サンプル）に対して適用し，後続の合成処理における話者特徴推定の安定化を図る．

2.2 ハイブリッド音声の生成

クローン音声と人工音声を混合比率 r（$0 \leqq r \leqq 1$）で線形加重し，ハイブリッド音声を作成した. 混合の定義を式（3）に示す．

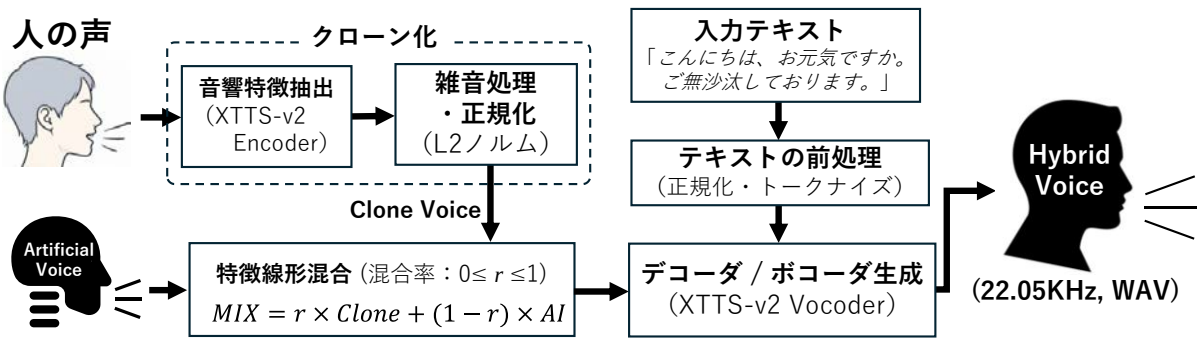

図 2 ハイブリッド音声の生成プロセス

$$MIX = r \times Clone + (1 - r) \times AI \quad (3)$$

混合に先立ち，各音声は同一デバイス上で生成し，振幅レベルのばらつきを抑えるために L2 ノルムに基づく正規化を施した．処理フローの概略を図 2 に示す．

本研究の混合は，聴取者が知覚しやすい「話者同一性（特定話者らしさ）」と「人工音声らしさ」の連続的な調整を意図している．線形加重は解釈が容易であり，r を操作変数として印象評価との関係を分析しやすい利点がある．また，刺激間での音量差が印象評価に与える影響を抑えるため，混合前の正規化によりレベル差をできる限り除去した．

2.3 刺激の出力と提示条件の統制

混合後の刺激は同一条件で WAV 化し，提示順序は参加者内でランダム化することで順序効果を低減した．また，言語条件間での比較可能性を高めるため，刺激生成に用いるモデル・計算環境・出力形式を固定し，音声処理（参照音声の前処理および混合）の手順も統一した．以上により，混合比率と（聴取者－話者）親密度が印象形成に与える効果を検討可能な刺激設計を構成した．

図 3 実験画面

3. 実験

中国語母語話者 10 名，日本語母語話者 9 名を対象として，作成したハイブリッド音声を聴取する評価実験を実施した．なお，いずれの参加者も音声サンプルの話者と面識があり，知人関係にある．

3.1 実験環境

参加者は静かな室内で，自身の PC に接続したヘッドホンを用いてオンラインで参加した．混合比率 r は 7 水準（1，0.8，0.6，0.5，0.4，0.2，0）とし，（$r = 1$）の条件は特定話者のクローン音声，（$r = 0$）の条件は人工音声の該当する．同一の意味を持つ一般的な挨拶文を刺激文章として用い，中国語は「你好，今天天气不错，好久不见」，日本語は「こんにちは，お元気ですか，ご無沙汰しております」，とした．

各言語につき，7 水準に基づく 7 種類の音声刺激（WAV，22.05 kHz）を作成した．提示は Python と Flask で実装した Web アプリケーションにより行い，各音声の長さは約 5 秒，試行間間隔は約 1 秒とした．実験終了後，同アプリケーションにより回答データは CSV 形式で自動的に保存することができる．

3.2 実験条件

参加者は自分の母語条件のみを聴取した．装着デバイスはイヤホンまたはヘッドホンとし，静穏環境での実施を求めた．

3.3 手続き

オンライン上で研究説明を提示し，同意を得た後に実験を開始した．参加者は各試行で音声（約 5 秒）を聴取しながら，「聞き取りやすさ」「好感」「親近感」「違和感」「よそよそしさ」の 5 項目を 5 件法で回答するように依頼した（図 3）．全 7 試行を連続して実施し，話者カテゴリおよび混合比率は明示しないブラインド条件とし，提示順序は参加者内でランダム化した．

4. 結果と考察

4.1 中国語母語話者によるハイブリッド音声の印象評価

表 1〜5 に，中国語母語話者を対象としたハイブリッド音声の聴取印象の評価結果を示す．実験参加者 10 名のうち，cn_7 は他の参加者と比較して条件間の評定変動が大きく，極端な値を示す傾向が見られたため，全体結果への影響を考慮して除外した．

その結果，各混合比率 r における聴取印象について，いずれの評価項目においても統計的に有意な差は確認されなかった（one-way repeated measures ANOVA，$F(6, 48) = 0.59$，$p > .1$）．

表 1　各混合比率のハイブリッド音声の印象評価
（Q1：聞き取りやすさ，中国語母語話者）

Q1	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
cn_1	5	5	5	5	5	5	5
cn_2	4	4	5	4	4	4	4
cn_3	5	5	5	5	5	5	5
cn_4	5	5	5	5	5	4	5
cn_5	4	4	4	4	4	4	5
cn_6	5	5	5	4	5	4	4
cn_8	5	5	5	5	5	5	5
cn_9	4	4	3	4	4	4	4
cn_10	5	5	5	5	4	5	5
Mean	4.67	4.67	4.67	4.56	4.56	4.44	4.67
S.D.	0.50	0.50	0.71	0.53	0.53	0.53	0.50

表 2　各混合比率のハイブリッド音声の印象評価
（Q2：好感，中国語母語話者）

Q2	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
cn_1	3	4	4	5	3	4	5
cn_2	3	4	5	4	4	2	3
cn_3	5	5	5	5	5	5	5
cn_4	4	4	3	4	3	4	4
cn_5	3	4	3	5	4	4	5
cn_6	4	4	2	3	3	2	4
cn_8	4	5	5	5	5	4	5
cn_9	3	2	3	2	3	2	2
cn_10	2	2	2	3	2	3	3
Mean	3.44	3.78	3.56	4.00	3.56	3.33	4.00
S.D.	0.88	1.09	1.24	1.12	1.01	1.12	1.12

表 3　各混合比率のハイブリッド音声の印象評価
（Q3：親近感，中国語母語話者）

Q3	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
cn_1	4	4	4	4	3	4	4
cn_2	3	5	4	4	4	3	4
cn_3	5	5	5	5	4	5	3
cn_4	4	5	3	5	3	4	3
cn_5	4	4	3	3	4	3	4
cn_6	4	4	2	3	3	2	4
cn_8	4	5	5	5	5	4	5
cn_9	3	2	3	2	3	2	2
cn_10	2	4	2	3	2	3	3
Mean	3.67	4.22	3.44	3.78	3.44	3.33	3.56
S.D.	0.87	0.97	1.13	1.09	0.88	1.00	0.88

表 4　各混合比率のハイブリッド音声の印象評価
（Q4：違和感，中国語母語話者）

Q4	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
cn_1	2	2	1	2	2	2	2
cn_2	2	2	1	3	2	3	1
cn_3	1	4	2	1	2	1	1
cn_4	2	1	3	1	2	1	3
cn_5	4	4	4	4	3	5	4
cn_6	2	2	4	4	4	3	3
cn_8	1	1	1	1	1	2	1
cn_9	3	4	3	2	2	3	3
cn_10	4	4	5	3	4	3	1
Mean	2.33	2.67	2.67	2.33	2.44	2.56	2.11
S.D.	1.12	1.32	1.50	1.22	1.01	1.24	1.17

表 5　各混合比率のハイブリッド音声の印象評価
（Q5：よそよそしさ，中国語母語話者）

Q5	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
cn_1	2	2	2	2	3	2	2
cn_2	1	1	1	1	1	1	1
cn_3	1	1	3	1	2	1	3
cn_4	2	2	2	1	2	1	3
cn_5	2	2	2	3	2	3	1
cn_6	2	2	4	4	3	4	3
cn_8	1	1	1	1	1	2	1
cn_9	3	4	3	2	2	3	2
cn_10	2	1	2	3	4	2	1
Mean	1.78	1.78	2.22	2.00	2.22	2.11	1.89
S.D.	0.67	0.97	0.97	1.12	0.97	1.05	0.93

表 6　各混合比率における中心化スコアの合計値
（中国語母語話者）

cn	混合比率						
質問	0	0.2	0.4	0.5	0.6	0.8	1
Q1	0.57	0.57	0.57	-0.43	-0.43	-1.43	0.57
Q2	-2.00	1.00	-1.00	3.00	-1.00	-3.00	3.00
Q3	0.29	5.29	-1.71	1.29	-1.71	-2.71	-0.71
Q4	-1.00	2.00	2.00	-1.00	0.00	1.00	-3.00
Q5	-2.00	-2.00	2.00	0.00	2.00	1.00	-1.00

そこで，補助的解析として，主観評価における個人差の影響を軽減した分析として，各参加者の評定値から当該参加者の全条件平均を減じた中心化スコアを算出し，その合計値の比較を行った（表 6）．この解析では，相対的な評価傾向を把握することを目的とした．なお，聴取者が抱く「親しみ」を評価する観点から，各質問の上位 2 値を網掛けした．

その結果，「好感」や「親近感」といった親しみに関係する項目（Q2，Q3）では，混合比率 $r = 0.5$ の条件が上位に位置する傾向が見られた．一方，「違和感」や「よそよそしさ」（Q4，Q5）では，$r = 0.5$ の条件は上位に含まれず，$r = 0$, $r = 0.5$, および $r = 1$ の条件で共通して低い値を示した．

これらの結果から，統計的な有意差は認められなかったものの，中国語母語話者においては，混合比率 $r = 0.5$ の条件で「違和感」や「よそよそしさ」が相対的に低く，一方で「好感」及び「親近感」は，高くなる傾向が見られた．このことから，$r = 0.5$ の条件におけるハイブリッド音声が，より「親しみ」のある声として知覚されやすい可能性が示唆された．

4.2　日本語母語話者によるハイブリッド音声の印象評価

表 7〜11 に，日本語母語話者を対象としたハイブリッド音声の聴取印象の評価結果を示す．実験参加者 9 名のうち，jp_8 は他の参加者と比較して条件間の評定変動が大きく，極端に高い評定を示したため，全体結果への影響を考慮して除外した．

表 7　各混合比率のハイブリッド音声の印象評価
（Q1：聞き取りやすさ，日本語母語話者）

Q1	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
jp_1	5	3	3	4	3	5	4
jp_2	4	4	3	4	3	3	2
jp_3	2	3	2	4	5	3	2
jp_4	4	4	4	4	4	4	4
jp_5	3	4	4	4	4	2	4
jp_6	2	3	3	2	2	2	2
jp_7	3	4	2	4	3	2	3
jp_9	3	4	4	2	3	4	3
Mean	3.25	3.63	3.13	3.50	3.38	3.13	3.00
S.D.	1.04	0.52	0.83	0.93	0.92	1.13	0.93

表 8　各混合比率のハイブリッド音声の印象評価
（Q2：好感，日本語母語話者）

Q2	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
jp_1	4	4	2	3	1	3	2
jp_2	4	4	3	4	4	3	3
jp_3	2	3	3	4	2	3	1
jp_4	3	2	3	4	3	3	3
jp_5	3	2	4	4	3	3	4
jp_6	4	4	3	2	2	2	2
jp_7	2	3	2	3	3	2	4
jp_9	4	2	2	2	1	3	1
Mean	3.25	3.00	2.75	3.25	2.38	2.75	2.50
S.D.	0.89	0.93	0.71	0.89	1.06	0.46	1.20

表 9　各混合比率のハイブリッド音声の印象評価
（Q3：親近感，日本語母語話者）

Q3	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
jp_1	2	2	3	2	1	4	2
jp_2	4	4	3	4	4	2	3
jp_3	2	2	2	3	1	2	1
jp_4	3	2	3	4	3	3	3
jp_5	2	4	4	4	3	3	3
jp_6	4	4	3	2	2	2	2
jp_7	3	3	2	3	3	2	3
jp_9	1	2	1	1	2	3	1
Mean	2.63	2.88	2.63	2.88	2.38	2.63	2.25
S.D.	1.06	0.99	0.92	1.13	1.06	0.74	0.89

表 10　各混合比率のハイブリッド音声の印象評価
（Q4：違和感，日本語母語話者）

Q4	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
jp_1	2	1	1	1	4	1	2
jp_2	1	4	4	1	1	2	4
jp_3	4	2	4	4	4	4	4
jp_4	2	4	2	2	4	2	2
jp_5	2	4	2	2	2	4	3
jp_6	2	2	2	3	3	3	4
jp_7	3	2	4	2	3	4	2
jp_9	4	3	4	4	3	2	5
Mean	2.50	2.75	2.88	2.38	3.00	2.75	3.25
S.D.	1.07	1.16	1.25	1.19	1.07	1.16	1.16

表 11　各混合比率のハイブリッド音声の印象評価
（Q5：よそよそしさ，日本語母語話者）

Q5	混合比率						
参加者	0	0.2	0.4	0.5	0.6	0.8	1
jp_1	1	1	4	2	4	2	2
jp_2	1	1	2	1	1	3	2
jp_3	4	2	4	2	3	5	4
jp_4	2	4	2	3	4	3	4
jp_5	3	2	3	2	2	2	2
jp_6	1	2	1	3	2	4	5
jp_7	4	2	4	4	3	4	2
jp_9	2	2	2	2	1	1	1
Mean	2.25	2.00	2.75	2.38	2.50	3.00	2.75
S.D.	1.28	0.93	1.16	0.92	1.20	1.31	1.39

その結果，各混合比率r における聴取印象についても，いずれの評価項目においても統計的に有意な差は確認されなかった（one-way repeated measures ANOVA，$F(6, 42) = 0.72$, $p > .1$）．

そこで，補助的解析として，中国語母語話者の結果と同様に，中心化スコアを用いて相対的な評価傾向を確認した（表 12）．聴取者が抱く「親しみ」を評価する観点から，各質問の上位 2 値を網掛けした．

表 12　各混合比率における中心化スコアの合計値
（日本語母語話者）

jp	混合比率						
質問	0	0.2	0.4	0.5	0.6	0.8	1
Q1	-0.29	2.71	-1.29	1.71	0.71	-1.29	-2.29
Q2	3.29	1.29	-0.71	3.29	-3.71	-0.71	-2.71
Q3	0.14	2.14	0.14	2.14	-1.86	0.14	-2.86
Q4	-2.29	-0.29	0.71	-3.29	1.71	-0.29	3.71
Q5	-2.14	-4.14	1.86	-1.14	-0.14	3.86	1.86

その結果，「好感」や「親近感」（Q2，Q3）において，混合比率$r = 0.5$ の条件が上位に含まれる傾向が見られた．一方，「違和感」や「よそよそしさ」（Q4，Q5）では，$r = 0.5$ の条件は上位に含まれず，$r = 0$，$r = 0.2$，および $r = 0.5$ の条件で共通して低い値を示した．これらの結果から，統計的な有意差は認められなかったものの，日本語母語話者においても，混合比率 $r = 0.5$ の条件で「違和感」や「よそよそしさ」が相対的に低く，一方で「好感」及び「親近感」は，高くなる傾向が見られた．このことは，日本語においても，$r = 0.5$ の条件におけるハイブリッド音声は，より「親しみ」のある声として知覚されやすい可能性が示唆している．

4.3　中国語母語話者および日本語母語話者に共通する印象形成の傾向

以上の結果を総合すると，中国語母語話者および日本語母語話者のいずれの群においても，クローン音声（$r = 1$）や人工音声（$r = 0$）の条件と比べ，両者を中庸に混合した $r = 0.5$ の条件が最も「親しみ」のある声として受け取られていたことが示唆される．ただし，これらの知見は限られたサンプルに基づくものであり，今後は参加者数の拡大や発話内容の多様化を図り，より統計的に確かな検証を行う必要がある．

5.　おわりに

本研究では，関係性の形成に寄与する「親しみ」のある合成音声の生成手法の構築を目的として，XTTS-v2 を用いて生成した特定話者の音声と人工音声を任意の比率で混合できるハイブリッド音声生成システムを新たに開発した．そして，このシステムにより，知人の音声と人工音声を様々な比率で作成したハイブリッド音声を用い，その聴取印象の特性を検討した．

その結果，知人の音声と人工音声の混合のバランスが，聴取印象の形成に一定の影響を及ぼすことが示された．特

に，統計的に有意な差は認められなかったものの，中国語母語話者および日本語母語話者のいずれの群においても，混合比率 $r = 0.5$ の条件で「違和感」や「よそよそしさ」が相対的に少なく，「好感」および「親近感」が高まる傾向が見られた．これらの結果から，言語を問わず，$r = 0.5$ の条件におけるハイブリッド音声が，より「親しみ」のある声として知覚されやすい可能性が示唆された．

本研究の知見は，対話型エージェントや感情共感型 AI における音声設計において，ユーザ特性や利用文脈に応じた音声スタイルの調整を行う際の有用な指針となり得る．

今後は，より多くの参加者を対象としたデータ収集と統計的検証を進めるとともに，話者の種類を拡張し，聴取者との関係性（有名人・非知人・家族など）が印象形成に及ぼす影響についても検討を行う予定である．最終的には，得られた知見を音声エージェントの設計に反映し，自然で親しみやすく，かつ文脈に応じて柔軟に適応する音声提示の実現を目指す．

参考文献

[1] Coqui AI. "XTTS-v2: New Version of the Open-Source Text-to-Speech Model." *Hugging Face Repository*, 2024. Retrieved from https://huggingface.co/coqui/XTTS-v2.

[2] Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. "The Influence of Synthetic Voice on the Evaluation of a Virtual Character." *Interspeech Conference Proceedings*, 2017, pp. 229–233. DOI: 10.21437/Interspeech.2017-325.

[3] Bruder, C., Breda, P., & Larrouy-Maestri, P. "Attractive Synthetic Voices." *Computers in Human Behavior: Artificial Humans*, 2025, 100211. DOI: 10.1016/j.chbah.2025.100211.

[4] Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., & Schuller, B. "The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech." *Unpublished Manuscript / Conference Paper*, 2018. DOI: 10.21437/Interspeech.2018-1093.

[5] Ephraim, Y., & Malah, D. "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985, vol. 33, no. 2, pp. 443–445. DOI:10.1109/TASSP.1985.1164550.

[6] Cohen, I., & Berdugo, B. "Speech Enhancement for Non-Stationary Noise Environments." Signal Processing, 2001, vol. 81, no. 11, pp. 2403–2418. DOI:10.1016/S0165-1684(01)00128-1.

[7] Cohen, I. "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging." *IEEE Transactions on Speech and Audio Processing*, 2003, vol. 11, no. 5, pp. 466–475. DOI: 10.1109/TSA.2003.811544.