情報処理学会論文誌 Vol. 49 No. 12 3908–3917 (Dec. 2008)

個人の音声を反映する映像エンタテインメントシステム

足立吉広[†1,†2] 大谷大和[†1] 川本真一[†1]
四倉達夫[†1] 森島繁生[†2] 中村哲[†1]

視聴者の顔を CG で再現し，CG キャラクタとして映画に登場させる Future Cast System（FCS）を改良し，視聴者の声の特徴をそのキャラクタの台詞音声へ反映させ，キャラクタの顔と声の一致度を向上させて音声を出力するシステムを構築する．あらかじめ構築した話者データベースから視聴者の知覚的類似話者を選出し，その話者の台詞音声を視聴者のキャラクタに割り当て，短時間で台詞音声を映像と同期出力するシステムを提案する．知覚的類似話者は，個人性の知覚と関係があると報告されている 8 つの音響特徴量による距離の線形結合を用いて推定する．声優による 60 種類の声質の台詞音声データベースを用いた音声出力同期システムを構築し，視聴者のキャラクタの顔と選択された音声の一致度に関して 5 段階の主観評価を行った．登場者数と話者データベースの規模，および類似話者の許容度の関係を予備実験により調査し，実験条件にあてはめたところ，予想される許容度約 51%に対して主観実験値において 35%の許容が確認され，全体として予備実験で得られた予想値の 68%が達成できた．

Visual Entertainment System Considering Personal Voice

Yoshihiro Adachi,[†1,†2] Yamato Ohtani,[†1]
Shinichi Kawamoto,[†1] Tatsuo Yotsukura,[†1]
Shigeo Morishima[†2] and Satoshi Nakamura[†1]

In this paper, we propose an improved Future Cast System (FCS) that enables anyone to be a movie star with own individuality in voice as well as faces. Previous system created a CG character which closely resembles the face of the audience; however the voice of the character was selected only considering gender. Therefore, the voice of a CG character is not enough to identify oneself from others. The proposed system produces much closer voice to the audience by selecting one from a voice actor database, where voice similarity of speaker is estimated using a combined feature of 8 acoustic features. After assigning one CG character to the audience, the system produces voices in synchronization with the CG character's movement. We constructed the speech synchronization system using voice actor database with 60 voice quality, and conducted the subjective evaluation experiments of voice similarity in five-grades. Achievement rate of the proposal method for theoretical figure that considered the allowance rate of selected speaker to the database size is 68%.

1. はじめに

1.1 研究背景

Future Cast System（FCS）[1] は，2005 年日本国際博覧会において公開された，誰もが簡単に映画に登場することができる世界初のエンタテインメントシステムである．FCS は視聴者の顔画像の撮影，顔形状の計測，Computer Graphics（CG）によるキャラクタ化，映画への登場をすべて自動で行う（**図 1**）．映画上の視聴者の CG キャラクタは，表情豊かに会話や演技をする．これまでの映画などの映像作品は，視聴者に一方的に提示されているだけであったが，FCS は自分自身や知人，友人が登場するため，視聴者が作品に関与する双方向のエンタテインメントであるといえる．

しかしながら FCS で反映される視聴者の特徴は，顔のみであった．音声や身体形状，動作などには個人の特徴は考慮されていないため，顔以外との特徴の不一致があった．

1.2 従来 FCS の分析

視聴者全員について，自分に対応するキャラクタが映画の中に登場する従来 FCS の体験直後に，視聴者に対し以下の項目についてアンケートを実施した．

(1) ご自身の顔が認識できましたか？
(2) ご自身の特徴を映像で見分ける際に注目する部位 6 項目（動作，体格，髪型，音声，表情，顔）について 3 段階で評価してください（3：つねに強く注目する，2：ときどき注目する，1：特に注目しない）
(3) キャラクタが話す台詞の声に違和感を感じますか？(5：とても感じる，4：感じる，3：どちらでもない，2：あまり感じない，1：感じない)
(4) 台詞の声のどの部分に違和感を感じましたか？(声，話し方，その他)

†1 国際電気通信基礎技術研究所音声言語コミュニケーション研究所
Advanced Telecommunications Research Institute International Spoken Language Communication Research Laboratories
†2 早稲田大学
Waseda University

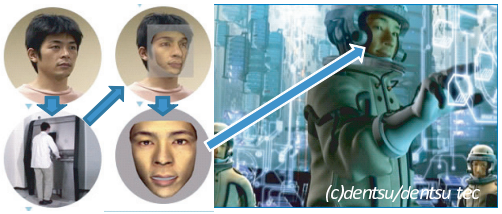

図 1　Future Cast System
Fig. 1　Future Cast System.

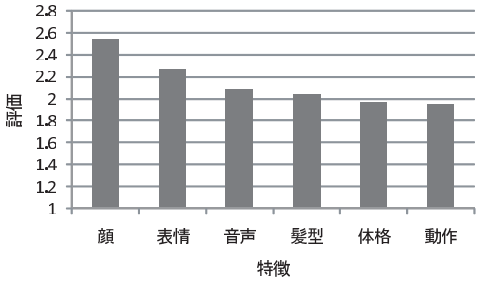

図 2　視聴者自身を見分ける際の各特徴の注目度
Fig. 2　Attention to each feature for distinction of oneself from others.

質問 (2) は FCS における個人性の特徴としての音声の注目度の調査であり，質問 (3)，(4) は音声の違和感の調査である．視聴者は 551 名（男性 295 名，女性 256 名，10〜71 歳）であった．

質問 (1) の結果では，自分に対応するキャラクタの顔が認識できた視聴者は 551 名中，328 名であった．以下，質問 (2)〜(4) の回答はこの 328 名による回答である．視聴者自身を見分ける際の特徴の着目点（質問 (2)）の評価結果を**図 2** に示す．グラフは 3 段階評価の平均値を表している．この結果から，音声は顔，表情の次に注目されており，個人を識別する特徴として 3 番目に重要であることが分かる．台詞音声の違和感（質問 (3)）に関する評価結果を**図 3** に示す．違和感を「とても感じる」，「感じる」と回答した被験者は全体の 65%である．音声の違和感のある部分（質問 (4)）の結果を**表 1** に示す．この結果から声質が主に違和感の原因となっていることが分かる．話し方も約 16%という結果から考慮する必要がある．これは年齢によるいい回しや，方言による韻律の変化などに対応することで改善されると考えられる．以上により，違和感のない本人に近い音声を提示する必要性が高

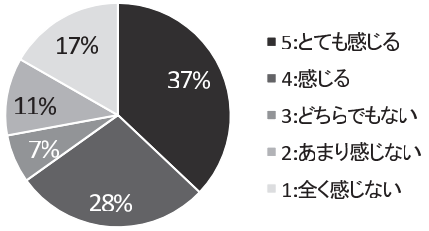

図 3　台詞音声の違和感
Fig. 3　Sense of mismatch in speech.

表 1　違和感の要素
Table 1　Component of the sense of mismatch.

	回答数	割合 [%]
声	239	71.1
話し方	53	15.8
その他	44	13.1

いことが分かる．

1.3　提案手法と関連研究

本研究では，声の違和感に対する改善法の条件として以下をあげる．

(1)　エンタテインメントとしてのクオリティ維持のため，音質と品質を保証

(2)　従来 FCS で用意されていた上映までの準備時間 15 分内に処理を完了

ここでいう音質の保証とは，ノイズなどにより台詞音声が劣化していないことを意味しており，品質の保証とは，演技力が十分であるということを示す．話者変換の技術で，違和感の改善を行う方法は，条件 (1) にあてはまらない可能性がある．この件に関しては，本論文中で実験と検証を行う．また，視聴前に本人がすべての台詞音声を収録することは，話者性という意味では自分のキャラクタには最適であるが，条件 (1)，(2) において疑問である．そこで以上の条件を考慮し，あらかじめ収録した複数の話者の音声から視聴者に似ている話者を選択し，その話者の音声を映画の中の自分のキャラクタに割り当てることで，声の違和感を改善する．そのためには視聴者の類似話者を話者データベースから選択する手法と，選択対象となる話者データベースの構築，さらに選択した音声をキャラクタに割り当て，映像と同期して視聴者に提示する仕組みが必要である．

これまで音声の話者性を扱った関連研究として，話者認識があげられる．話者認識は，申

告話者本人を検索する課題であり，音声から構築される Gaussian Mixture Model（GMM）などのモデルにおける類似度を用いていた[5]．本研究では知覚的に類似している他話者を検索することが目的であり，話者認識での手法が効果的であるかは未知である．音声の知覚的な類似度と音響特徴量の関係について，Amino らは音節を対象として，ケプストラム距離と知覚的類似度に強い相関があると報告している[2]．また，永嶋らは発話速度とイントネーションをできる限り同一に発声した音声を対象として，2–10 kHz のスペクトルと知覚的類似度に相関関係があると報告している[3]．本研究では，個人の特徴が表出されやすいイントネーションや発話速度なども含めて類似度を推定するために，複数の特徴を考慮した知覚的類似度の推定法[4] を用いる．

本論文では以上の実現方法を提案し，提案システムの有効性について検証を行った結果について報告する．2 章では提案する音声出力システムの概要について述べ，3 章では話者データベースの構築について述べる．4 章では知覚的類似話者の推定方法について述べ，5 章ではシステムの実装について述べる．6 章では提案システムの評価について述べ，7 章では本システムにおける議論を行う．8 章では結論と今後の課題について述べる．

2. 音声出力システム概要

FCS の従来システムと提案システムの音声出力方法を**図 4** に示す．従来システムは計測した視聴者の顔形状に対し男女判定を行い，判定結果に基づいて音声トラックの性別を決定する．音声トラックはキャラクタごとに男女各 1 トラックのみであるため，被験者の年齢や方言などの個人性は反映されない．音声トラックが決定したら，音楽（BGM: Background Music）と効果音（SE: Sound Effect）の音声トラックと同期して出力する．

提案システムは，視聴者に類似した話者を話者データベースから選択し，視聴者の個人性を反映した音声をキャラクタに割り当て，映像と同期して出力する仕組みが必要である．図 4 に示す提案システムでは，事前準備としてすべての話者がすべてのキャラクタの台詞音声を収録した話者データベースを構築しておく．システム利用時には，最初に視聴者から音声を収録する．この音声を用いてデータベースから，最も類似する音声の話者を選択する．選択された話者の音声から，視聴者に割り当てられたキャラクタに必要な音声のみを「加工」へ送る．加工は将来的に音質劣化のない韻律・声質制御などの適用を可能にする．しかし現在の技術では音質劣化が懸念されることから，本研究では変換は行わない．この後音声は再生リストへ追加される．この再生リストは，台詞音声ファイルの再生のタイミング情報が書かれている．FCS 上映時には BGM，SE と，絶対時間情報を表す LTC（longitudinal

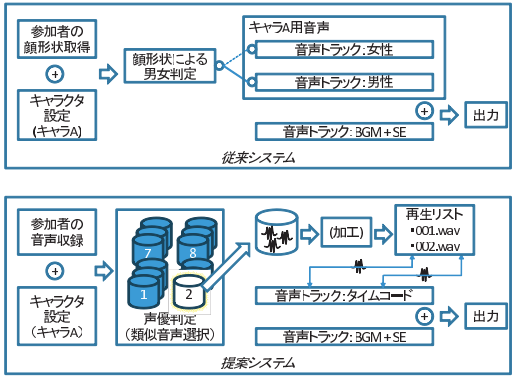

図 4　FCS の音声出力方法
Fig. 4　Sound output system of FCS.

time code）信号を再生する．この LTC 信号が再生リストに書かれた時間に達したときに台詞音声が再生される．これにより台詞ごとに映像と同期することが可能である．再生リストを用いることにより，上映と並行して音声処理が可能であり，かつ発話単位での分散処理が可能となる．これにより上映までの準備時間を短縮することができる．

以上の提案システムを構築するために話者データベースの構築と，入力音声の類似話者を選択する手法が必要である．以下の章ではこれらについて説明する．

3. 話者データベースの構築

提案システムにおける話者データベースの構築は，作品の品質に深く関係するため重要である．映画のキャラクタの動きに合わせて演技した音声を収録することは，声優などの経験者以外の人には難しい．そこで声優による収録により台詞音声データベースを構築する．収録内容は，2005 年日本国際博覧会において公開された FCS 用の映像作品 “Grand Odyssey” の全 89 台詞である．1 名の声優につき 2 種類のキャラクタを想定し演技してもらい，30 名の声優から 60 種類の音声（合計 5,340 サンプル）を収録し，データベースを構築した．

4. 知覚的類似話者の選択

4.1 知覚的類似度の推定式

知覚的類似話者を選択するためには，1 種類の音響特徴量より複数の音響特徴量を用いた方が，精度良く推定できると報告されている[4]．そこでこの文献 4) の手法に従い，個人性が関係する音響特徴量の距離を線形結合し，知覚的類似度を推定する．なお，類似話者選択精度に関しては文献 4) で検討が行われている．

知覚的類似度 s の推定式を式 (1) で表す．

$$s = -\sum_{i=1}^{n} \alpha_i x_i \tag{1}$$

n は結合する音響特徴量の数，x_i は i 番目の音響特徴量における距離，α_i は線形結合の係数である．

4.2 音響特徴量

音響特徴量の距離算出には，個人性の知覚に関係があると報告されている 8 種類の音響特徴量を用いた．以下で説明する音響特徴量はすべて，分析窓長 25 ms，フレーム周期 10 ms で求める．

MFCC　MFCC（Mel Frequency Cepstral Coefficient）は雑音環境に頑健であり，音声認識や話者認識に用いられている[5]．本研究の MFCC は，MFCC12 次元，ΔMFCC12 次元，Δ パワー 1 次元の合計 25 次元の特徴量ベクトルを表す．

STRAIGHT Cepstrums　北村らは高次ケプストラムに現れる声帯音源の情報と声帯音源の周波数特性の傾斜が，個人性知覚に影響を与えることを示している[6]．そこでこのときの音響特徴量である対数 STRAIGHT[7] スペクトルをフーリエ変換して求めたケプストラムの 35 次以上と，そのケプストラムの 1 次を扱う．

スペクトル　高次のスペクトルもまた音声の個人性と強い関係がある[3]．そこで報告された 2.6 kHz 以上の高次スペクトルを扱う．

STRAIGHT-Ap　齊藤らは STRAIGHT の分析パラメータである非周期性指標 STRAIGHT-Ap が，個人性の知覚に影響があることを示した[8]．そこで 2 kHz 以下の帯域の STRAIGHT-Ap を扱う．

基本周波数　橋本らは，基本周波数は個人性に影響があることを示している[9]．そこで基本周波数（F0）も音響特徴量の 1 つとして扱う．基本周波数は STRA-IGHT の分析の一部である STRAIGHT-Tempo によって抽出する．

フォルマントとスペクトル傾斜　声質は音声の類似度の判定に主要な音響特徴量である．木戸らはフォルマントとスペクトル傾斜は声質の表現に必要な特徴量であるとしている[10),11)]．そこで 1 次から 4 次のフォルマントと 3 kHz 以下の対数スペクトルの傾斜を扱う．

4.3 距 離 尺 度

音響特徴量の距離尺度には，人による類似度判定を考慮し，イントネーションのような大局的な音響特徴量の時間変化を扱う DTW 距離[12)] を用いる．特徴ベクトル間の距離はユークリッド距離とした．また，DTW 距離は発話時間（時系列長）によって正規化した．

4.4 結合係数 α_i の最適化

式 (1) により得られる知覚的類似度の推定値と，人による知覚的類似度の相関が高くなるように，結合係数 α_i の最適化を行う．最適化には女性話者 36 名が「あらゆる現実をすべて自分の方へねじまげたのだ」と発話した，文献 4) と同一の音声を用いた．また知覚的類似度は，類似度の評価基準を一定に保つため，話者と面識のない正常聴力を有する 20 代男性 1 名によって与えられた．最適化方法は，まず話者データベースから 1 名のターゲット話者を選択する全パターンに対し，データベース内の全音声を人手によりクイックソートのアルゴリズムで類似度順に並べる．次に順列で表現された知覚的類似度の順位を，最も精度良く再現できるように結合係数 α_i を，最急降下法により求める．クイックソートでの類似度の判定には，声質やイントネーション，発話速度といった 1 つの特徴に注目せず，全体的な印象によって類似度の判断を行った．複数の音声ペアの類似度を相対的に比較することは，単一の音声ペアに絶対値で類似度を付与するより再現性が高いと考えられる．知覚的類似度による順位とその推定順位の相関は，式 (2) で示される Spearman の順位相関係数 ρ で評価した．

$$\rho = 1 - \frac{6\sum_{i=1}^{N}(a_i - b_i)^2}{N^3 - N} \tag{2}$$

a は被験者によって並べられた知覚的類似度の順列における順位，b は音響特徴量によって並べられた順列における順位，N は順列の長さを示す．本実験では順列の長さ N は 36 である．

5. システム実装

提案システムの実装を**図 5** に示す．

(1)　被験者は Recording PC + Headset を用いて話者選択用の音声を収録する．

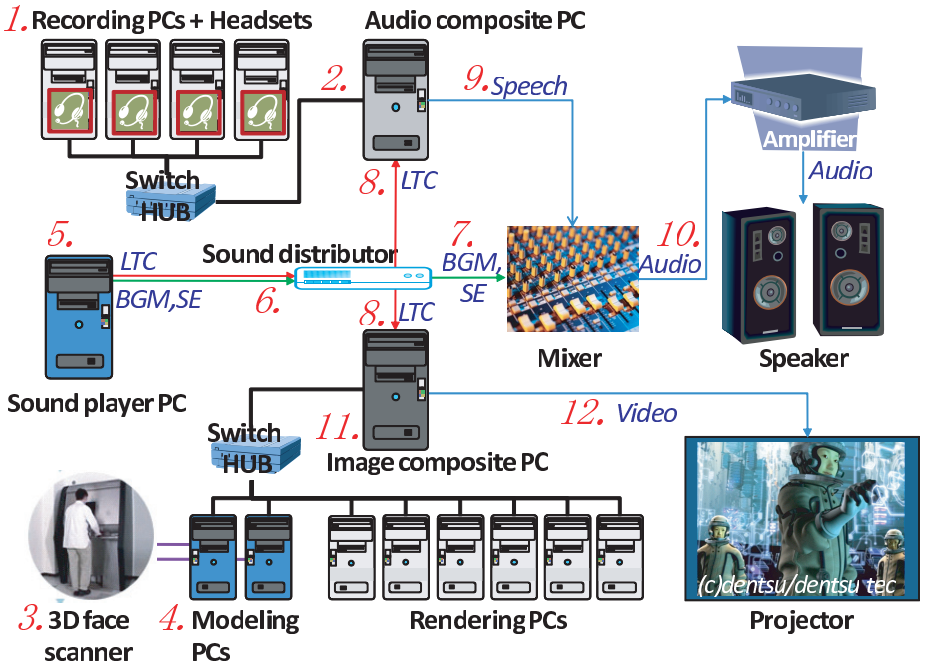

図 5　システムの実装
Fig. 5　System implementation.

(2)　収録した音声と最も知覚的類似度の高い話者名を，被験者名とともに Audio composite PC に送る．
(3)　一方 3D 顔モデルの構築のために 3D face scanner で被験者の顔形状を取得．
(4)　Modeling PCs において 3 次元顔モデルを構築．
(5)　Sound player PC から LTC と BGM，SE のステレオ信号を音声出力する．
(6)　このステレオ信号は Sound distributor に入力．
(7)　BGM，SE はそのまま Mixer へと送られる．
(8)　Sound distributor は LTC を分配し Audio composite PC と Image composite PC に送る．
(9)　LTC を受け取った Audio composite PC は，所定の時刻に発話用音声を Mixer へ送り，BGM，SE とミキシングを行う．

図 6　類似話者選択用音声の収録
Fig. 6　Recording for selecting similar speaker to a visitor from speaker database.

(10)　Mixer の音声出力は Amplifier で出力を調節され Speaker へ送られる．
(11)　一方，LTC を受けった Image composite PC は，プリレンダリングされている背景と，被験者の 3D 顔モデルを合成した画像を Rendering PC から読み込む．
(12)　所定の時刻に Projector へ映像を出力する．

図 6 にシステム運用時に実際被験者が音声を収録している様子を示す．

6.　提案システムの評価

提案システムの評価として，声質変換法との比較評価，FCS へ適用した際の有効性に関する評価，そして性能評価を行う．

6.1　声質変換法との比較

視聴者の声の個性をキャラクタに反映させる手法として，データベースから知覚的に類似する声優を選択するほかに，声質変換の技術を用いてデータベースの音声の話者性を変換する方法が考えられる．そこでこれらの比較評価を行う．声質変換法として，音声合成手法 STRAIGHT に基づいた高音質な声質変換手法[13] を検討する．なお，この声質変換手法による音声の音質評価と変換精度に関しては，文献 13) で検証が行われている．

構築した 60 種類の声質を含むデータベースから選択した知覚的類似話者の音声と，知覚的類似話者上位 3 名の音声を用いて入力話者に似せた声質変換音声を比較する．評価基準は，「FCS として好ましい音声について」である．実験手順は，最初に被験者に入力話者の音声，類似話者選択による音声，声質変換による合成音声を提示し，次の質問を行う．質問内容は「映画としては，合成音声より，自分に似た声優さんの音声を使う方が良いと思いますか？」である．これに対し被験者は 5 段階（5：とてもそう思う，4：そう思う，3：どち

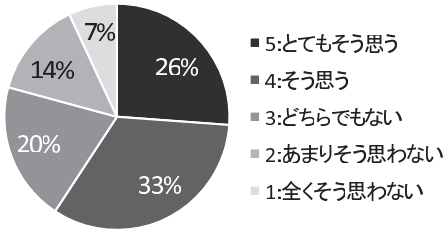

図 7　声質変換法と類似話者選択法の比較実験結果（質問内容「映画としては，合成音声より，自分に似た声優さんの音声を使う方が良いと思いますか？」）
Fig. 7　Comparative estimate of the voice conversion and the similar speaker selection.

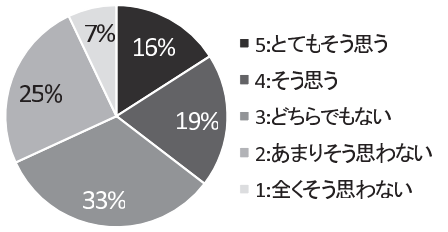

図 8　顔に対する声の一致感（質問内容「声優の音声は自分に合っていましたか？」）
Fig. 8　Degree of voice correspondence for character's face.

らでもない，2：あまりそう思わない，1：まったくそう思わない）で回答する．

実験に用いた文章は，“Grand Odyssey” の「さすが天才プログラマー」という台詞である．データベースの音声は，サンプリング周波数 48 kHz，量子化ビット数 16 bit で収録されている．被験者は 2008 年 2 月に東京の日本科学未来館において FCS を一般公開した際に参加した 130 名である．

実験結果を**図 7** に示す．本実験では音声を聞き比べたうえで，選択されただけの音声が合成音声より好まれたことが分かる．合成音声は入力話者に合わせて声質を変換できる利点があり，スペクトル距離においては選択音声より似ている．一方話者選択では，あらかじめ収録している声質しか選べないが，音質や品質は良い．本結果の主な原因は音質であると考えられる．被験者に提示した合成音声と声優による音声は，音質に差があることは明らかであった．これは文献 13) で検証されているとおりである．視聴者が，音質が十分でない合成音声より声優の音声を選ぶことは自然なことであり，実験結果は妥当であると考えられる．FCS は映画のような作品性を求めて作られている．そのためこのような作品では，音質劣化は極力避けるべき問題であり，類似話者の選択は有効な手法であると考えられる．

6.2　提案システムの有効性

提案手法を組み込んだ FCS を被験者に体験してもらい，その直後に次の質問を行う．

(1)　ご自分の顔を認識することができましたか？
(2)　声優の音声は自分に合っていましたか？
(3)　声の質が自分とまったく違う声優の声を出す今までのシステムより，自分に似ている声優の音声を出す方がおもしろいと思いますか？

被験者は (1) に対しては「はい・いいえ」で回答し，他は 5 段階（5：とてもそう思う，4：そう思う，3：どちらでもない，2：あまりそう思わない，1：まったくそう思わない）で回答する．

被験者は 6.1 節と同日，同会場で実験に参加した 172 名である．上映内容は “Grand Odyssey” である．最大同時上映参加人数は 20 名であるが，実験では 15 名で行った．

有効回答数は 144 であり，質問 1 で自分自身の顔を認識できたと回答した人数は 113 名であった．この 113 名の質問 2 に対する回答を**図 8** に示す．5 段階評価の平均値は 3.12 であり，自分に合っていたと感じる評価が多かったことが分かる．類似話者を選択したにもかかわらず顔に合っていたと感じた結果が十分ではない原因として，データベースの規模，選択対象が声優の音声であること，本人自身を聞きなれていないなどが考えられる．話者データベースの規模として 60 種類の声質は，老若男女，様々な方言の視聴者を考慮した場合，十分な量ではないことが考えられる．これについては 6.3 節で議論する．また，声優は演技力が視聴者自身とかけ離れているため，「そんなに上手く演じたところを見たことがない」といった違和感が生じてしまった可能性がある．しかしそのような違和感を解消するために，演技力が十分でない素人の音声をデータベースに含めると，映画としてのクオリティの低下が懸念されるため，演技力と類似性はトレードオフの関係が予想される．さらに，視聴者が自分自身の音声を聞きなれておらず，選択結果の音声が骨導音を除いた気導音のみの音声であることも原因の 1 つと考えられる．

次に質問 3 に対する回答を**図 9** に示す．76%の被験者が自分に似ている声優を割り当てられることが面白いと感じていることが分かる．一方，残りの 10〜24%の被験者は，自分に似ている音声を好まないことが分かった．被験者の中には，スピーカからの聞きなれない自分の声を好ましくないと思っている人や，そもそも自分の声に対するコンプレックスを持つ人もいる．実験結果から，提案システムの精度が向上したときに面白いと感じさせられる

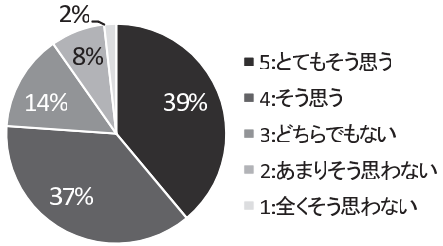

図 9　類似声優割当ての期待（質問内容「声の質が自分とまったく違う声優の声を出す今までのシステムより，自分に似ている声優の音声を出す方がおもしろいと思いますか？」）
Fig. 9　Expectation of appropriating a similar speaker.

視聴者は，およそ 76%であることが分かった．

6.3　データベース規模に対する有効性の検討

6.2 節で，選択声優の音声が自分に合っていた（評価 4 以上）との回答は約 35%であった．しかし選択される類似話者は，データベースの規模にも依存すると考えられる．そこでさらに本実験結果を検証するために，データベースの規模とその中で選択可能な類似話者に対する違和感を調べる実験を行った．

検討に用いた音声は，成人男女 28 名の声優による音声である．発話内容は母音をバランス良く含んだ 1 文として「あめんぼあかいなあいうえお」を用いた．実験の被験者は正常な聴力を有する 20 名（男性 8 名，女性 12 名，19〜24 歳）である．評価方法は，1 対の音声を話者 A，話者 B の順に聞かせ，次の質問に数値で回答させた．質問内容は「話者 A の役の声を話者 B に変えたときどのくらい違和感を感じますか？」であり，回答は 0〜100 で答える（目安は 0：とても違和感を感じる，70：許容できる，100：まったく違和感を感じない）．データセットはカウンタバランスをとっており，話者 A，B が同一話者のデータに関しては，評価値 100 が得られるものと仮定して実験から除外した．20 名分の評価値の平均を，その評価対のスコアとした．これらの評価値から必要なデータベースの規模を推測する．

Leave-one-out cross validation を実施し，28 名中 1 名を類似話者選択システムの入力を想定して選んだ入力想定話者，残り 27 名をデータベース用の音声とする．この 27 名から無作為に N 名に絞り込み，この中で入力想定話者として最も違和感のない話者を求める．すべての組合せについて算出し，28 通りの入力想定話者の平均を求めた．この評価値を，入力想定話者数に対し N 倍のサイズのデータベースから類似話者を選択する際に評価される予想値とした．$N = 1$〜27 としたときの評価値を**図 10** に示す．横軸を対数としたとき，$N = 10$

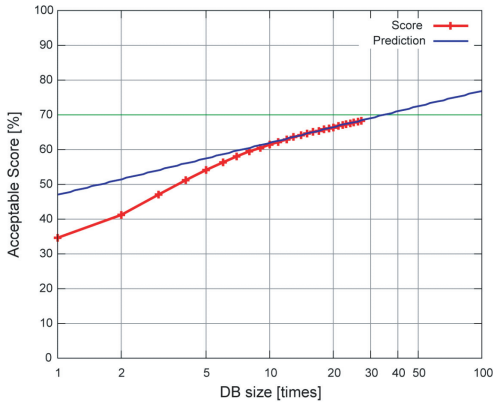

図 10　“Grand Odyssey”（FCS）に必要なデータベースの規模
Fig. 10　Necessary database size for “Grand Odyssey” (FCS).

以上からほぼ線形であることから，$N = 20$ の傾きを基準に線形式をあてはめると，「70：許容できる」となるのは $N = 30$〜40 である．実験に用いた音声が成人男女の音声であるため，仮に年齢層を子供・成人・高齢者の 3 クラスに分けることができるとした場合，高年齢者・子供も同規模のデータベースが必要とすると，全体では 90〜120 倍のデータベースが必要となる．このことより 20 名参加型の “Grand Odyssey” の FCS で Acceptable Score を 70%にするためには，1,800〜2,400 名の話者のデータベースが必要である試算となる．

本研究で用意したデータベースの規模は 60 種類であり，同時参加人数が 15 名であることから，DB size は 4 [times] となる．この場合の Acceptable Score（理論値）は 51.27%であり，選択された声優の音声が自分に合っていた（図 8，評価 4 以上）との回答約 35%は，理論値の約 68%にあたる．100%に達しない理由は 6.2 節に示したことが考えられる．

6.4　話者選択を適用した FCS の音声出力システムの評価

提案音声出力システムにおける，話者選択のための音声収録から上映準備完了までの所要時間について評価を行った．被験者から話者選択用の音声として，「あらゆる現実をすべて自分の方へねじまげたのだ」という文章の音声を収録する．この音声を FCS に入力する．FCS 内部での処理は，入力音声に対して話者データベース内のすべての声優の同一発話内容の音声との類似度を算出し，最も類似度の高い声優の台詞音声をキャラクタの台詞音声として割り当てる．この内部処理の必要時間を算出する．

入力話者は，20 代から 40 代の被験者 13 名である．本システムの動作環境は，Microsoft

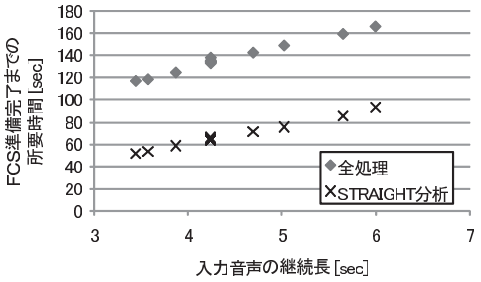

図 11　音声出力の準備所要時間
Fig. 11　Processing time for sound output.

Windows XP，Intel Core2 CPU T7600，2.33 GHz × 2，3.00 GB RAM である．

入力音声の継続長と準備完了までの所要時間の結果を**図 11** に示す．入力音声の分析には STRAIGHT を用いており，この部分の処理時間も同図に示す．準備時間の平均所要時間は 137.09 秒，最大でも 165.89 秒であり，10 分ほどの時間を音声収録に用いることができる．本実験結果は，1 文の発話を用いた結果であるため，10 分以内に 1 文の音声を収録すればよく，20 名同時参加の FCS に対しても，複数台の計算機で並列処理をすれば可能であるといえる．以上から FCS の準備時間の条件 15 分以内に処理できることが確認できる．

所要時間から入力音声の音響特徴量を分析する時間を除いた約 60 秒の間では，入力話者と話者データベース内の全話者の対に対し距離計算を行っている．つまり全探索である．もし話者データベースに 1,800 名いる場合は 1,800 秒かかる見込みとなる．提案システムが 20 名専用ではないため，計算時間は重要な問題である．要求時間を満たすためには，複数の PC を用いた分散処理，もしくは効率良く類似話者を検索する手法が必要である．効率的な検索方法として，データベース内の話者をクラスタリングし階層構造にしておくことで，距離計算の回数を減らすなどが考えられる．

7. ディスカッション

7.1 データベースの規模の妥当性

提案システムでは，音声のクオリティを維持するために 20 名規模，許容度 70%で 1,800 名の声優データベースが必要であるとの試算を示した．しかしながら，これほど多くの声優を対象に，台詞音声を収録，準備することは妥当な方法であるとはいい難い．

声優の人数を抑える方法として，話者変換技術がある．話者変換技術では，1 人の音声から様々な声質の音声を生成できるが，問題は音質劣化である．しかし STRAIGHT による合成音声は自然な聴覚的印象であり[14]，音声の発話速度や基本周波数を制御して，話声を歌声に変換する研究[15] や，声質と歌い回しを転写する研究[16] が報告されている．また話者の特徴であるイントネーションに関しても，音質劣化の少ない制御法が提案されている[17]．このような音質を劣化させずに声質を変化させる技術は，声質の種類を増やすために有効な手段であり，今後の課題である．

7.2 データベースの声質分布

声質を表現するための要素は検討されている[18] が，声質の分布の検証は非常に難しい研究課題であるといえる．今回用意した 60 種類の音声は，視聴者は 9 歳から 60 歳程度で男女同数程度を想定し収録されている．予備実験として 60 種類の音声に対し，話者が何歳に聞こえるか，どちらの性別に聞こえるか主観評価実験を行ったところ，8〜63 歳に聞こえ，性別の判定もほぼ同数であるという結果を得られた．このことから，話者の年齢と性別という項目において，今回用いた声優による話者データベースの声質分布は妥当であると考えられる．当然声質には上記以外にも特徴があると考えられるが，評価軸となる特徴が確立しておらず未検討である．そのため，これらに関しての分布が妥当であるかを検証することは今後の課題である．また，声質を表現する特徴量の一覧が得られたとして，分布が一様になるよう意図した声質の話者を探し，収録するのは困難であると考えられる．そのため，声質のバランスを考慮した収録法もまた今後の課題となる．

声優の音声は，発声法などのトレーニングなどにより，発話方法，声質において視聴者の発話とは分布が異なることが考えられる．発話方法の違いに関しては，演技のための発話と日常での発話に差があって当然であるので，分布の違いは問題ない．声質に関しては声優には少ない声質の視聴者がいることも考えられる．この場合，視聴者に類似話者として割り当てた音声が似ていないこともありうる．こうした場合は，

(1)　選択用のデータベースに一般人を含めて，視聴者の分布に近いものにする，

(2)　声質変換技術により声質を近づける，

などの対策があげられる．対策 (1) の場合は，話者データベースの演技力の低下が課題となる．対策 (2) は，音声モーフィングを想定している．話者間のモーフィングによって声質の内挿は可能である[19] が，外挿は品質的に難しいと考えられる．そのため，できるだけ稀な話者が必要となり，そのような話者を効率的に探す方法も課題といえる．

8. ま と め

FCS において，話者データベースから視聴者の類似話者を選択し，その話者の音声をシーンに同期して再生することで，視聴者の声の個性をキャラクタに反映させ，キャラクタの顔との不一致を改善するシステムを提案した．類似話者の選択を採用した提案システムは，声質変換技術を用いる場合と比較したところ，59%の被験者に支持された．また 60 種類の声質のデータベースを用いた提案システムに対し，顔と声の一致度に関する評価を 5 段階の主観評価を行ったところ，データベースのサイズに対する類似音声の許容の理論値を，68%達成した．さらに提案した音声出力システムは，視聴者の音声の入力から FCS の上映準備を完了までの処理時間の要求条件を満たしている．

今後の課題として類似話者選択手法の音響特徴量の最適化があげられる．類似話者の選択精度は，提案システムにおいて重要であり，さらには人間が知覚的に類似していると感じる尺度を明らかにすることは，類似度の定量化の観点から望まれることである．提案システムで用いた類似話者の選択法では，個人性に関係があると報告されている複数の音響特徴量を用いたが，このうちどの特徴量が重要であるかは現在検討中である．

謝辞　本研究は文部科学省の科学技術振興調整費の助成を受けた．

参 考 文 献

1) Morishima, S., Maejima, A., Wemler, S., Machida, T. and Takebayashi, M.: Future Cast System, *ACM SIGGRAPH 2005 Sketch*, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc 2, ISBN 1-59593-099-X, 020-morishima.pdf (2005).
2) Amino, K., Sugawara, T. and Arai, T.: Speaker Similarities in Human Perception and their Spectral Properties, *Proc. WESPAC IX* (2006).
3) 永嶋育美，高際光晴，齋藤　裕，柳川博文，長尾優次，村上　尚，福島　学：人の話者識別における音声の類似性の検討，音響学会春季講演論文集，3-P-5, pp.737–738 (2003).
4) Adachi, Y., Kawamoto, S., Morishima, S. and Nakamura, S.: Perceptual Similarity Measurement of Speech by Combination of Acoustic Features, *Proc. ICASSP*, pp.4861–4864 (2008).
5) Reynolds, D.A.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE Trans. Acoust. Speech and Audio Processing*, Vol.3, No.1 (1995).
6) 北村達也，斎藤　毅：単母音の個人性知覚における各種音響特徴量の寄与，日本音響学会講演論文集，pp.443–444 (2007).
7) Kawahara, H.: STRAIGHT: An extremely high-quality VOCODER for auditory and speech perception research, *Computational Models of Auditory Function*, Greenberg and Slaney (Eds.), pp.343–354, IOS Press (2001).
8) 齋藤　毅，北村達也：3 連続母音に含まれる個人性情報の知覚的要因，日本音響学会講演論文集，pp.441–442 (2007).
9) 橋本　誠，樋口宜男：音声の個人性知覚における既知話者未知話者の影響，音響講論（秋），pp.263–264 (1996).
10) 箕輪有希子，木戸　博，粕谷英樹：声質表現語の音響関連量—予備的検討，日本音響学会講演論文集，pp.363–364 (2000).
11) 木戸　博，箕輪有希子，粕谷英樹：声質表現語の音響関連量に関する非線形分析—決定木による方法，日本音響学会誌，Vol.58, No.9, pp.586–588 (2002).
12) 迫江博昭，千葉成美：動的計画法を利用した音声の時間正規化に基づく連続音声認識，日本音響学会誌，Vol.27, No.9, pp.483–490 (1971).
13) 大谷大和，川本真一，戸田智基，中村　哲，鹿野清宏：STRAIGHT モーフィングに基づく特定話者音声の生成，日本音響学会講演論文集，1-11-29 (2008).
14) Matsui, H. and Kawahara, H.: Investigation of Emotionally Morphed Speech Perception and its Structure using a High Quality Speech Manipulation System, *Proc. Eurospeech'03*, pp.2113–2116 (2003).
15) 齋藤　毅，後藤真孝，鵜木祐史，赤木正人：SingBySpeaking：話声を歌声に変換する歌声合成システム，日本音響学会講演論文集，1-11-28, pp.305–308 (2008).
16) 河原英紀，生駒太一，森勢将雅，高橋　徹，豊田健一，片寄晴弘：モーフィングに基く歌唱デザインインタフェースの提案と初期的検討，情報処理学会インタラクションの理解とデザイン特集号，Vol.48, No.12, pp.3637–3648 (2007).
17) 足立吉広，森島繁生：話者のイントネーションを模倣するインタラクティブ声質変換システムの構築，インタラクション 2005 論文集，Vol.2005, No.4, pp.261–268 (2005).
18) 木戸　博，川股正佳，粕谷英樹：声質評価システムの構築，電子情報通信学会技術研究報告，Vol.103, No.252, pp.1–6 (2003).
19) Kawahara, H. and Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation, *Proc. ICASSP*, pp.256–259 (2003).

(平成 20 年 3 月 24 日受付)
(平成 20 年 9 月 10 日採録)

足立　吉広

2005 年成蹊大学大学院博士前期課程修了．現在，早稲田大学大学院博士後期課程在学中．2006 年 9 月〜2009 年 3 月（株）国際電気通信基礎技術研究所（ATR）音声言語コミュニケーション研究所研修研究員．音声の韻律変換に関する研究，音声の知覚的類似度の推定に関する研究に従事．日本音響学会学生会員．

大谷　大和

2005 年大阪大学基礎工学部システム科学科卒業．2007 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学博士後期課程在学中．主に音声合成の研究に従事．日本音響学会，電子情報通信学会，ISCA 各学生会員．

川本　真一（正会員）

1998 年九州工業大学情報工学部電子情報工学科卒業．2000 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2005 年同大学博士後期課程修了．博士（情報科学）．同年（株）国際電気通信基礎技術研究所（ATR）音声言語コミュニケーション研究所に入社，現在に至る．音声情報処理，マルチモーダル情報処理の研究に従事．電子情報通信学会，日本音響学会各会員．

四倉　達夫

1998 年成蹊大学工学部電気電子工学科卒業．2000 年同大学大学院修士課程修了．2000〜2001 年（株）ATR 知能映像通信研究所研修研究員．2003 年成蹊大学大学院博士課程修了．博士（工学）．同年（株）国際電気通信基礎技術研究所（ATR）音声言語コミュニケーション研究所に入社，現在に至る．デジタルコンテンツ制作支援技術，コンピュータグラフィックス，顔モデリング・アニメーションに関する研究に従事．2000 年電子情報通信学会学術奨励賞，同年 NICOGRAPH/MULTIMEDIA 論文コンテスト最優秀論文賞．ACM，電子情報通信学会，画像電子学会各会員．

森島　繁生（正会員）

1987 年東京大学大学院電子工学専門博士課程修了，工学博士．同年成蹊大学工学部電気工学科専任講師，1988 年同助教授，2001 年同電気電子工学科教授．2004 年早稲田大学理工学部応用物理学科教授，現在に至る．1994〜1995 年トロント大学コンピュータサイエンス学部客員教授，1999 年より国際電気通信基礎技術研究所客員研究員を併任．現在，明治大学非常勤講師，新潟大学非常勤講師，早稲田大学デジタルエンタテインメント研究所所長．コンピュータグラフィックス，コンピュータビジョン，音声情報処理，ヒューマンコンピュータインタラクション，感性情報処理の研究に従事．1991 年本会業績賞受賞，顔学会理事．IEEE，ACM，日本音響学会，映像情報メディア学会，日本心理学会各会員．

中村　　哲（正会員）

1981 年京都工芸繊維大学工芸学部電子工学科卒業．1981〜1994 年シャープ（株）勤務．1992 年京都大学博士（工学）．1994〜2000 年奈良先端科学技術大学院大学助教授．2000 年より（株）国際電気通信基礎技術研究所（ATR）．現在，音声言語コミュニケーション研究所長，および，（独）情報通信研究機構上席研究員，音声言語 GL．独カールスルーエ大学客員教授，けいはんな連携大学院教授．音声翻訳，音声認識等の音声言語情報処理の研究に従事．電気通信普及財団賞，情報処理学会山下賞，AAMT 長尾賞，ドコモモバイルサイエンス賞，情報処理学会業績賞，日本音響学会技術開発賞受賞．IEEE，電子情報通信学会，日本音響学会各会員．