情報処理学会論文誌 Vol. 51 No. 2 250–264 (Feb. 2010)

来場者の声の特徴を反映する映像エンタテインメントシステムのための台詞音声生成システム

川本真一[†1,†2] 足立吉広[†2,†3] 大谷大和[†2]
四倉達夫[†2] 森島繁生[†1,†3] 中村哲[†1,†2]

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，視聴者から収録した少量の音声サンプルを用いて，視聴者に似た台詞音声を生成するため複数手法を統合し，生成された台詞音声をシーンに合わせて同期再生することで，視聴者の声の特徴をキャラクタに反映させるシステムを提案する．話者データベースから視聴者と声が似た話者を選択する手法（類似話者選択技術）と，複数話者音声を混合することで視聴者の声に似た音声を生成する手法（音声モーフィング技術）を組み合わせたシステムを構築し，複数処理を並列化することで，上映準備時間の要求条件を満たした．実環境を想定してBGM/SEを重畳した音声によって，従来手法である類似話者選択技術より得られる音声と，提案法で導入した音声モーフィング技術より得られる音声を主観評価実験により評価した結果，Preference Scoreで56.5%のモーフィング音声が目標話者の音声に似ていると判断され，音声モーフィングを組み合わせることでシステムが出力する台詞音声の話者類似性を改善できることを示した．

Voice Output System Considering Personal Voice for Instant Casting Movie

Shin-ichi Kawamoto,[†1,†2] Yoshihiro Adachi,[†2,†3]
Yamato Ohtani,[†2] Tatsuo Yotsukura,[†2]
Shigeo Morishima[†1,†3] and Satoshi Nakamura[†1,†2]

In this paper, we propose an improved Future Cast System (FCS) that enables anyone to be a movie star while retaining their individuality in terms of how they look and how they sound. The proposed system produces voices that are significantly matched to their targets by integrating the results of multiple methods: similar speaker selection and voice morphing. 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 a voice actor database with 60 different kinds of voices. Our system achieved higher voice similarity than conventional systems; the preference score of our system was 56.5% over other conventional systems.

1. はじめに

Future Cast System (FCS)[1] は，2005 年日本国際博覧会において初めて公開された，誰もが簡単に映画に登場することの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・顔形状を取り込み，視聴者の CG キャラクタを作成，映画への出演を短時間で実現することができる．視聴者の顔の特徴を反映させたキャラクタが即座に映画に登場し，表情豊かに演技を行う新しい視聴者参加型エンタテインメントシステムとして注目されている．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．

FCS の拡張として足立ら[2] は，視聴者の声の特徴を映画のキャラクタの台詞音声へ反映させ，登場キャラクタの顔と声の一致度を向上させるシステムを提案した．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，登場キャラクタに割り当てることで，登場キャラクタの顔と声の一致度の高い台詞音声を同期出力する．演技の品質や音質・話者性については，話者データベースに依存し，短い処理時間で安定した台詞音声を出力できるという利点がある．一方で，十分な話者性を確保するためには話者データベース構築に多くの時間・コストを必要とする．関連する研究としては，音声合成や声質変換の研究はさまざま行われており[3)-5]，人間の声と聞き比べても遜色のない自然性が確保されているものもある[6]．これらの技術は限られた話者データベースから多様な話者・発話内容の音声を生成する技術として多方面への応用が期待されるが，映像作品に使われた例はほとんどない．特に，エンターテインメントシステムとして音声合成や声質変換技術を運用するうえで，最終的な出力となる台詞音声の音質を保証することが重要である．

そこで本研究では，先行研究の短い処理時間で安定した音質を確保できる足立ら[2] のシステムを拡張し，限られた話者データベースを用いて，視聴者の声の特徴を映画のキャラク

†1 独立行政法人情報通信研究機構
National Institute of Information and Communications Technology
†2 株式会社国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International
†3 早稲田大学
Waseda University

タの台詞音声へ反映させるための複数手法の統合システムを提案する．視聴者から収録した少量の音声から，短時間で複数手法の処理を並列に実行し，得られた複数手法の結果を統合することで，提案システムの出力する台詞音声の話者類似性の改善を目指す．

2. システム設計

映像作品としては，1つの要因の品質低下が全体の品質に大きく影響するため，映像品質に見合う音質が確保されていることが必須である．また，映像と音のずれは視聴者に違和感を感じさせることがあるため，台詞音声は映像と同期して出力されるべきである．エンターテインメントシステムとしては，視聴者への負担が極力少ないことが重要である．特に音声収録作業に慣れていない視聴者にとっては，長時間の収録は負担となるため，収録時間や内容についても十分配慮する必要がある．また，対象とする作品が変わったときに，台詞数や登場人物数に対応できるシステムの柔軟性や，視聴者から収録した音声が少量であってもシステムを安定して運用できる頑健性も重要である．さらに，視聴者が作品を鑑賞するまでの待ち時間も極力短いことが望ましい．従来 FCS では上映までの準備時間が 15 分以内であり，本システムにおいてもこの条件を満たすべきである．FCS のように視聴者の誰もが簡単に映画に登場人物として参加できるエンターテインメントシステムを構築するうえで，視聴者の声の特徴を反映させた台詞音声の出力には，1) 映像作品に期待される音質を確保し，2) 視聴者の収録などの負担を極力軽減し，3) 短い待ち時間で，4) 映像に合わせて台詞音声を同期出力する，5) 頑健性・柔軟性を有するシステムであることが重要である．

本論文において，声の特徴の反映をより積極的に行うため，先行研究である足立ら[2] のシステムに新たな手法を追加し，結果を統合することで，話者類似性の改善を目指す．本章では，上記要求条件を満たすためのシステム拡張，および手法選択について論じる．

2.1 話者データベースの整備

話者データベースの整備は，出力する映像作品の品質に深く関係するため，重要である．特に，映像内のキャラクタの演技にマッチする台詞音声を発話することは，声優など経験者以外には困難である．本論文では，足立ら[2] のシステムに使用した話者データベースを基に，音声サンプルの分割位置を調整したものを用いた．本データベースは，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 台詞の台詞音声（110 音声サンプル），および類似話者選択用の 1 文章の読み上げ音声（1 音声サンプル）からなる 111 音声サンプルについて，1 名の声優につき 2 名の異なる話者を想定した声で演技してもらうことで，30 名の声優から 60 名分の仮想的な話者の音声（「1 名分の仮想的な

話者につき 111 音声サンプル」×「60 名分の仮想的な話者」= 合計 6,660 音声サンプル）を収録したものである．

類似話者選択用の文章については，個人の特徴を効率的に獲得するために，音素がバランス良く含まれていることが望ましい．特に音声の個人性は，子音よりも母音に多く現れるといわれているため，特に母音のバランスは重要と考える．さらに，想定される視聴者は老若男女さまざまであり，音声の収録対象文章としては，複雑な文章や長い文章は避け，簡単で誰もが発話しやすい文章であることが望ましい．そこで本システムでは日本語の母音をバランス良く含み，多くの人になじみのある文章「あめんぼあかいなあいうえお」を採用した．また，追加手法の音声の収録対象文章についても共通化することで，収録作業の軽減を図った．

映像内のキャラクタの口の動きが見てとれるようなシーンにおいては，口の動きと台詞音声との同期（リップシンク）が映像作品の自然性に大きく影響する．そのため，映像と背景音（BGM）・効果音（SE）との同期だけではなく，台詞音声との同期も実現する必要がある．そこで映像作品 “Grand Odyssey” の台詞のうち，口の動きが見てとれるシーンの台詞を中心にリップシンク実現のための音素および音素継続長情報を新たに整備した．整備手順は，まず整備対象となる台詞音声を音声認識ソフトウェアで音素アラインメントを行い，得られた音素および音素継続長よりテストアニメーションを生成した結果を目視で検査を行い，不具合のあったアニメーションの音素および音素継続長を修正した．本研究では，音声認識ソフトウェアとして ATR 音声言語コミュニケーション研究所で開発された ATRASR[7] を用いた．ATRASR は，認識処理において大人用音響モデルと子供用音響モデルの両方を用いたパラレルデコーディング[8] を行い，リップシンクアニメーションへの応用実績[9] のあるソフトウェアであることから，多様な声質を含む話者データベースの音素アラインメント処理に有効であると考えられる．キャラクタの口の動きが見てとれるようなシーンの発話 2,220 音声サンプルについて，発話アニメーションの目視検査でリップシンクの精度を確認し，セグメンテーション結果の修正を必要とした発話は 27 音声サンプル（1.2%）であった．

2.2 音声モーフィングに基づく新たな手法の導入

通常の映像作品では，台詞音声は収録スタジオで収録した音声に BGM やエフェクトを加えて出力するため，本論文で提案するシステムの出力する台詞音声も同様に高品質である必要がある．しかし，視聴者の音声を収録する環境は，必ずしも収録スタジオのように環境の整った場所ではなく，環境雑音が混入することもある．このような雑音環境下においても，混入した雑音が音質に影響しにくい手法であることが望ましい．

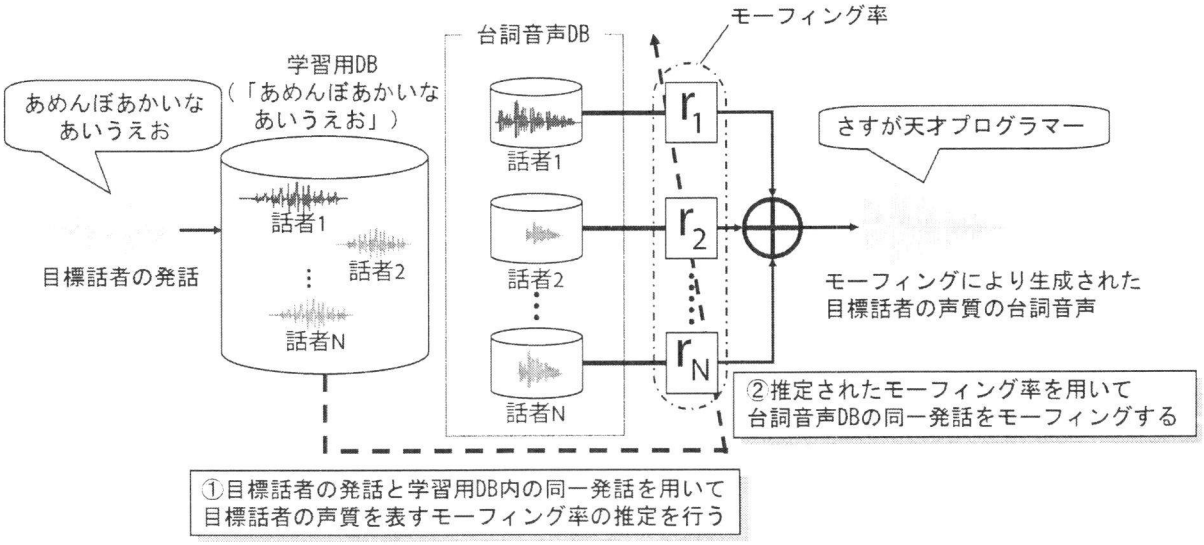


図 1　音声モーフィングに基づく台詞生成システム
Fig. 1　Dialogue generation system based on auditory morphing.

音声収録に慣れていない人にとっては，音声収録作業は負担となる可能性があるため，その作業は極力短時間にする必要がある．このため，視聴者から収録した少量の音声に対して安定して動作する手法であることが必要である．

また，生成される台詞音声に対しても，リップシンクが実現されるべきである．

つまり，視聴者から収録した少量の音声に対して安定して動作し，雑音環境下でも安定した音質を確保でき，リップシンクのための情報を効率的に取得できる手法が必要である．そこで本論文では，STRAIGHT に基づく音声モーフィング[10] による特定話者音声生成技術を新たに開発した．音声モーフィング[11] とは入力された 2 つの音声から得られるパラメータを任意の配分で混合することより，中間的な特徴を持った合成音声を生成する手法である．Kawahara らが提案した STRAIGHT[6] に基づく音声モーフィングでは，音声パラメータとして，STRAIGHT スペクトル，非周期性指標，$F0$，を用い，時間軸，および周波数軸とともに，STRAIGHT スペクトログラム上に付与した特徴点に基づいて共通のモーフィング率により 2 つの音声間のモーフィングを実現する．**図 1** に本システムにおける提案手法の概要図を示す．提案手法ではまず目標話者の音声から学習用 DB 内の音声に対するモーフィング率を推定する．次に推定されたモーフィング率に基づき台詞 DB 内の音声をモーフィングすることで目標話者の声質の台詞音声を生成する．提案手法の詳細は 4 章において述べる．本手法は以下のような特徴を有する．

安定した音質確保　高品質の分析合成系として定評のある STRAIGHT を基礎とする手法

であり，音質劣化の少ない台詞音声の出力が期待できる．また，視聴者の音声から推定するパラメータは音声モーフィングのための少数のモーフィング率のみであり，少量の音声からでも安定してモーフィング率を推定できる．さらに，事前に準備した雑音のない話者データベース内の音声からモーフィング音声を生成するため，直接的に音声スペクトルを操作する手法と比べ，視聴者の音声に雑音が混入している場合でも，出力される台詞音声は安定した音質を確保できる．

安定したリップシンク精度　話者データベースの台詞音声に対して事前に音素および音素継続長が付与されているため，音声モーフィングによって得られる台詞音声の音素および音素継続長は，重みつき線形和の計算により容易に算出することが可能である．また，事前に発話アニメーションの目視検査により整備された音素および音素継続長情報を基礎とする安定したリップシンク精度を実現できる．

2.3　処理の並列化

視聴者にとって映像を見るまでの待ち時間は，サービスの印象に大きく影響する．特に，安定したサービスとして提供することを考える場合，映像の上映開始は時間厳守であり，指定時間に確実に上映できることが必須である．このため，視聴者の音声収録時間などにより処理時間が変動する状況下でも，限られた時間内で得られる最良の結果を確実に提供するシステム設計が重要である．また，複数の視聴者の音声に対して，複数の手法を適用する場合は，必要とする処理が並行に処理されるよう，システムを設計することが重要である．さらに，処理に使用できる PC の台数の増加にともない，処理時間を短縮できるようなスケーラビリティのあるシステム設計であることが望ましい．

そこで，音声収録処理，音声分析・合成処理，音声出力処理が独立して動作するように設計した．音声分析・合成処理については，音声サンプルごと，手法ごとに複数の PC で並列に動作する独立したサーバとして処理単位を設計することで，PC の台数増加にともない処理時間が短縮できるようなシステム設計を行った．さらに音声分析処理を共通化することで，処理の効率化を行った．**図 2** に音声収録と音声分析合成システムの構成を示す．

2.4　映像同期音声出力

先行研究である Maejima らの従来システム 1（性別推定に基づく音声選択システム）[1]，足立らの従来システム 2（類似話者選択に基づく音声選択システム）[2] と提案システムの音声出力方法を**図 3** に示す．

従来システム 1 [1] では，事前準備としてすべての台詞に対して男女の標準音声を準備し，視聴者から取得した顔画像データから性別推定を行うことで，視聴者に割り当てられたキャ

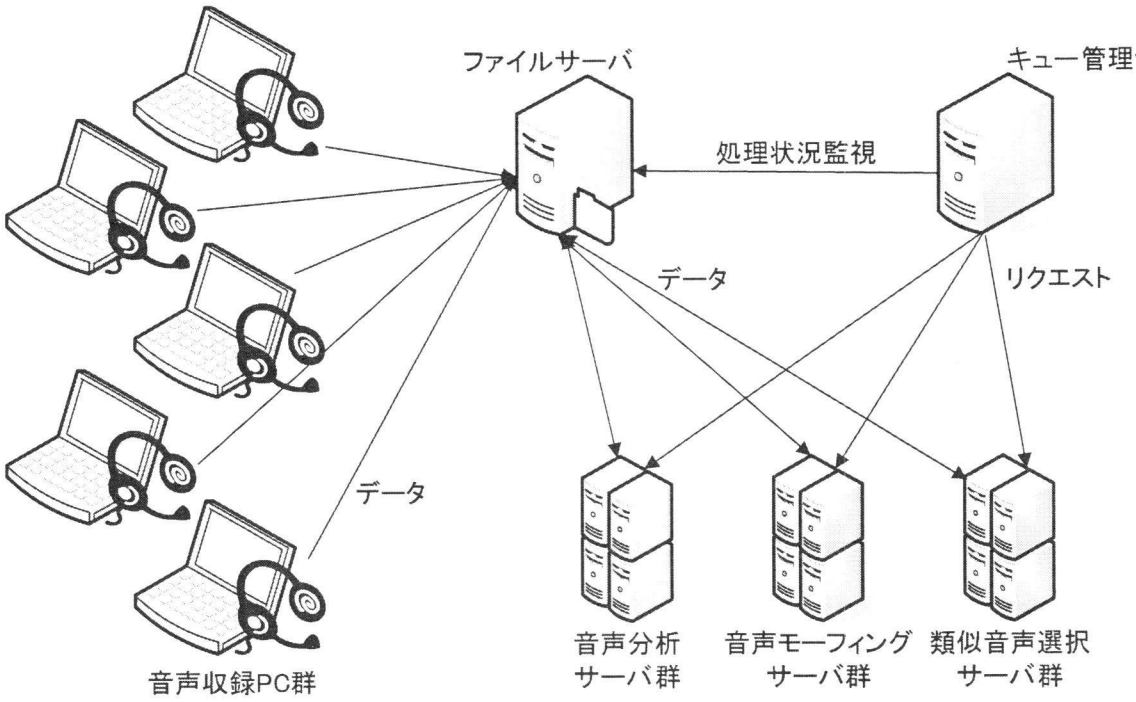


図 2　音声収録と音声分析合成システム
Fig. 2　Voice recording and analysis/synthesis system.

ラクタに男女どちらの音声を割り当てるかを決定する．従来システム 2[2] 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，および台詞音声ファイルの再生のタイミング情報が書かれた再生リストを準備しておく．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台詞音声のみを抽出する．映像上映時には BGM，SE と，絶対時間情報を表す LTC（longitudinal 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させ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．

提案システムでは，さらに従来システム 2[2] における類似話者選択処理と並行して，音声モーフィング処理を行う．また，従来システム 1[1] で用いていた顔画像による性別推定も行う．従来システム 1[1] における手法を，本システムの枠組みに統合するために，台詞音声を発話単位に切り出し，各発話のタイミングを統一するよう従来システム 1 の標準音声デー

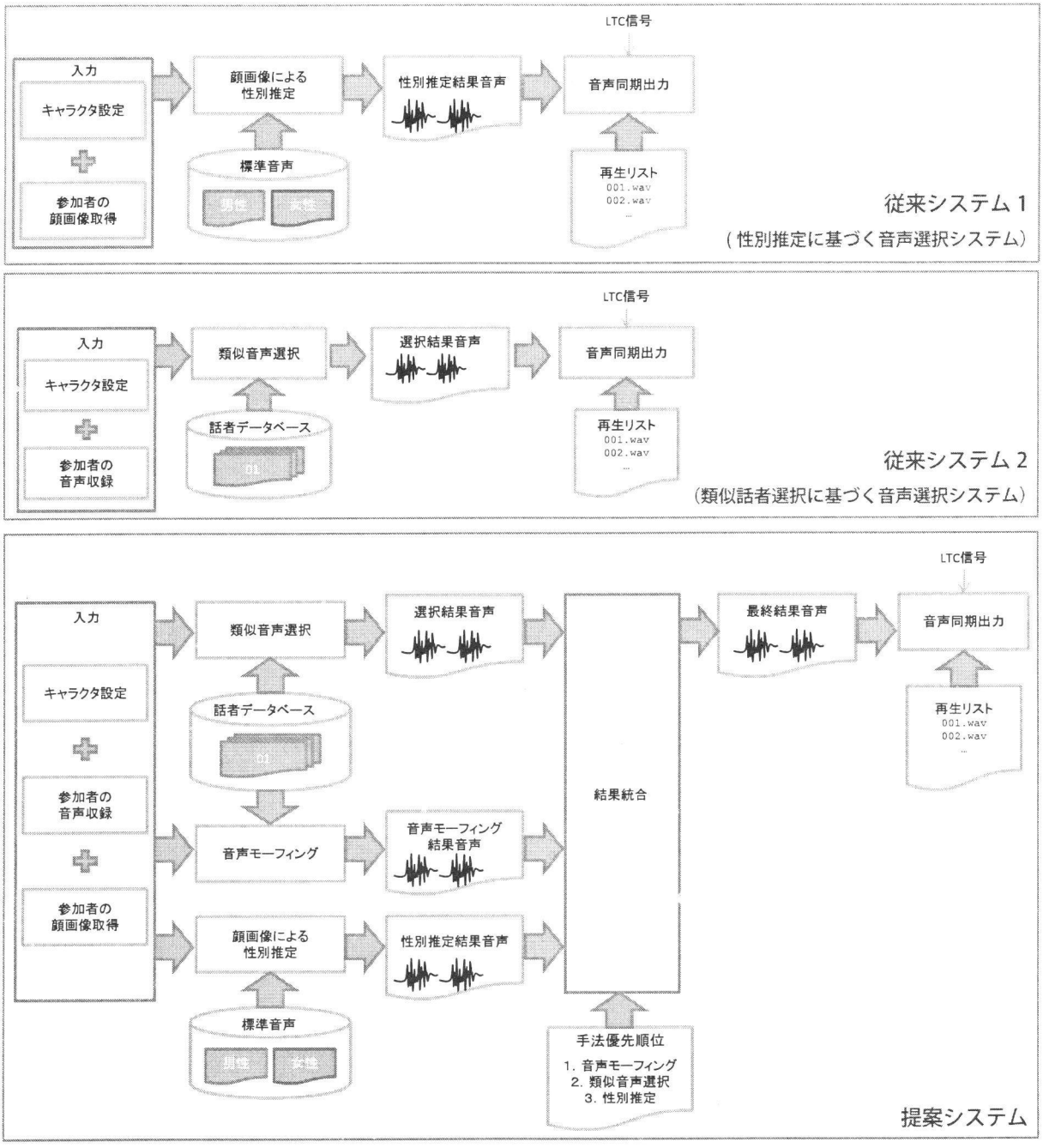


図 3 音声出力システムの比較
Fig. 3 Comparison of sound output systems.

タベースを修正した．並列に処理された類似話者選択，音声モーフィング，顔画像による性別推定の結果を，事前に設定した手法の優先順位をもとに統合し，最終的な台詞音声を決定する．最終的に決定された台詞音声に付随するリップシンクのための情報も結果統合時に確

定し，映像出力処理へ情報を提供する．

2.5　結果統合による処理の安定化

本論文では，事前に各手法の優先順位を設定し，以下の手順で結果統合を行う．

Step.1　各手法から得られた台詞音声を取得

Step.2　取得した各手法の台詞音声について音質を評価し，音質が十分でない台詞音声を破棄

Step.3　Step.2 の処理において破棄されなかった各台詞音声について

- **Step.3.1**　優先順位 1 位の手法の台詞音声が存在すれば，その手法の台詞音声を出力
- **Step.3.2**　優先順位 1 位の手法の台詞音声が存在せず，2 位の手法の台詞音声が存在すれば，優先順位 2 位の手法の台詞音声を出力
- **Step.3.3**　すべての手法の台詞音声が存在しなければ，あらかじめ設定した標準音声を出力

Step.1 は上映前の任意の時間において，現時点での処理結果を確定させるために実行する．上映直前の決められた時間に，その時点での最良な結果を最終的な音声として確定させることで，上映時間の安定化を図った．音声が収録されていない状態や処理結果が得られていない状態での上映に対応するため，標準音声を設定した．本論文では，Maejima らのシステム[1] で得られる台詞音声を標準音声として用いた．

本手法では Step.2 において音質に関する評価を実施することで，最終的に出力される台詞音声の音質を保証する．理想的には，音質評価も自動で行われることが望ましいが，現状では実運用に耐えうる確実な手法は見当たらない．そこで本論文では確実性を重視し，音声モーフィングに基づく新たな手法から得られる台詞音声については，人手で聴取により音質を確認した．この聴取検査をスムーズに行うために，手法ごとに出力先を分け，聴取検査対象音声を明示するようにシステムを設計した．なお，Maejima らのシステムで用いた手法（性別推定に基づく手法）[1]，足立らのシステムで用いた手法（類似話者選択に基づく手法）[2] はともに，視聴者にあった台詞音声を事前に整備したデータベースから選択する処理であり，結果として得られる台詞音声の音質劣化はないと判断し，Step.2 における音質評価処理を省いた．

本手法では Step.3 では各手法の優先順位を決定する必要がある．本論文では，Step.2 において最終的に出力される台詞音声の音質を保証することを前提としているため，出力音声の話者性を基準として優先順位を決定する．類似話者選択に基づく手法，および新規に追加した音声モーフィングに基づく手法に関する話者性の評価結果については，5.2 節で述べる．

提案システムの枠組みは非常にシンプルであり，容易に拡張可能である．音声モーフィング以外の手法についても，入力（視聴者の音声，キャラクタ情報など）と出力（発話ごとに切り出され，発話タイミングが統一された台詞音声一式）の仕様を一致させ，Step.3 における優先順位，Step.2 における音質評価方法を決定することで，他の手法も同様に統合することができる．

3. 知覚的類似話者の選択

3.1 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[12]．そこでこの文献 12) 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 12) 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．

3.2 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，サンプリング周波数 16 kHz，分析窓長 25 ms（400 点），フレーム周期 10 ms（160 点）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[13]．本研究の MFCC は，MFCC 12 次元，ΔMFCC 12 次元，Δ パワー 1 次元の合計 25 次元の特徴量ベクトルを用いる．

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[14]．そこでこのときの音響特徴量である対数 STRAIGHT[6] 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 次を用いる．

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[15]．そこで報告された 2.6 kHz 以上の高次スペクトルを用いる．

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[16]．そこで 2 kHz 以下の帯域の STRAIGHT-Ap を用いる．

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[17]．そこで基本周波数（F0）も音響特徴量の 1 つとして用いる．基本周波数は STRAIGHT 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[18),19)]．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用いる．なお本研究において，スペクトル傾斜は STRAIGHT-Spectrum の 0〜3 kHz に対する単回帰直線を最小二乗法により求め，その直線の傾きを用いた．また，フォルマントは 16 次の線形予測分析（LPC; Linear Predictive Coding）により得られる多項式の根を求め，その多項式の根に対応する周波数を用いた[20]．

3.3　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[21] 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．

3.4　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 12) 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 名の目標話者を抽出し，残りの 35 名を人手で目標話者との類似度順にクイックソートのアルゴリズムで並べる．この人手による類似度付与作業を 36 名から目標話者を選択するすべての組合せ（36 通り）について行う．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．人手による類似度付与の際に「話者 X と話者 A はどの程度似ているか」といった絶対的な類似度の付与より，「話者 X は話者 A と話者 B のどちらに似ているか」といった比較結果から相対的な類似度による順位を付与する方が高い再現性を

実現できると考え，類似度による順位を評価基準として用いた．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[22]．

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

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．

本手法の知覚的類似度に関する評価によると，人手により類似度の順列を 2 回再現する実験における Spearman の順位相関係数が 0.72 であるのに対し，本手法によって類似度の順列を推定したものと，人手により類似度の順列を与えたものとの Spearman の順位相関係数は 0.66 であった[12]．

4. 音声モーフィングによる特定話者音声生成

4.1 複数話者を用いた音声モーフィング

高橋らは従来の STRAIGHT モーフィングを拡張し，複数話者を用いたモーフィングの手法を提案している[23]．複数話者を用いたモーフィングの手順は従来の手法とほぼ同じである．**図 4** に 2 話者を用いた場合のモーフィングの手順を示す．まず，モーフィングに用いるすべての音声に対して，モーフィング時の音韻の時間情報およびフォルマント位置が一致するようにいくつかの特徴点を付与する．モーフィング率 r に基づき，モーフィング後の特徴点の位置を決定し，特徴点以外の点を特徴点間において区分線形補間を行うことで，時間軸および周波数軸を伸縮させる．そして，各特徴量を各話者に与えられたモーフィング率に基づいて次式のように複数の音声特徴量を線形結合することでモーフィング音声を得る．

$$\boldsymbol{x}_{mrp} = \sum_{s=1}^{S} r_s \boldsymbol{x}_s \tag{3}$$

ここで，r_s は s 番目の話者に付与されるモーフィング率，$\boldsymbol{x}_s$ は s 番目の話者の時間周波数に対して伸縮した特徴量を表す．

4.2 STRAIGHT モーフィングによる特定話者音声の生成

高橋らが提案した複数話者の音声を用いた STRAIGHT モーフィング[23] は，話者間の中間的な音声を生成することを目的とする手法であり，手動で任意のモーフィング率を与えることでさまざまな声質を容易に生成できる．したがって，ある一定のモーフィング率を与え

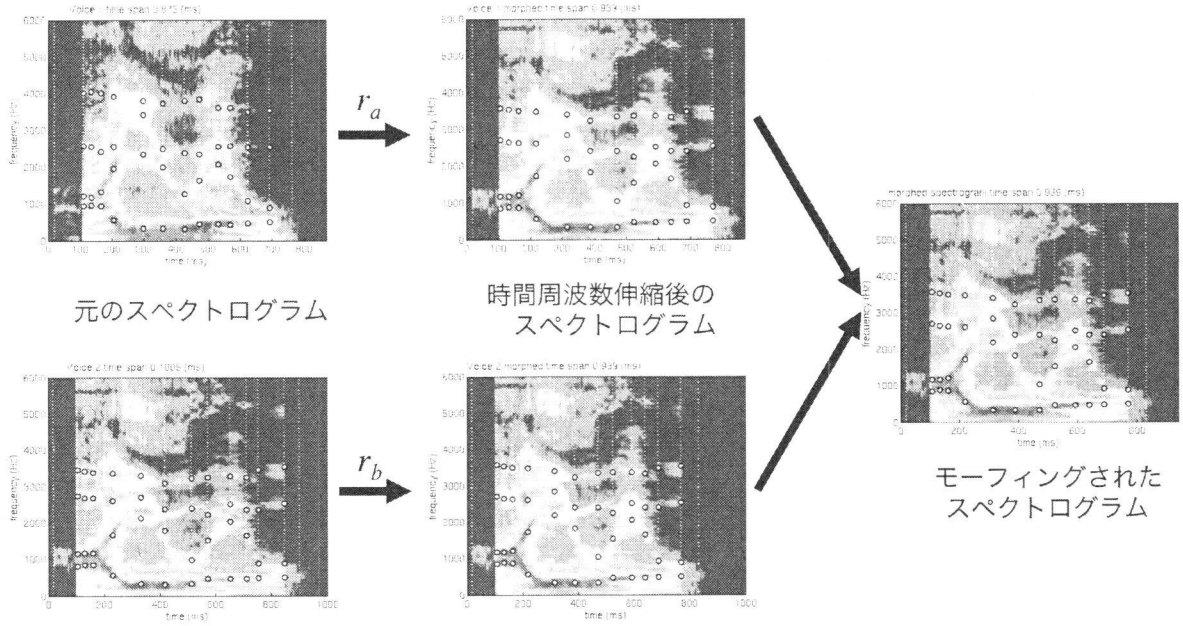


図 4　2 話者を用いた音声モーフィング
Fig. 4　Auditory morphing using 2 speakers.

ることで，特定話者の声質を生成することが可能であると考えられる．しかし，従来法では手動によりモーフィング率を決定するため，特定話者の声質を表すモーフィング率を設定することは困難である．

筆者らは，話者データベースから複数話者音声を抽出し，それらの音声を混合することで，話者データベースに含まれない新たな話者の音声を生成することを目的とし，音声の混合手法としての高橋らの手法を拡張し，特定話者音声生成を複数話者の STRAIGHT モーフィングにおけるモーフィング率最適化問題として解く手法を提案する．提案手法では，ある発話で推定したモーフィング率を別の発話に適用することで，データベースに事前に準備したさまざまな発話文章に対する特定話者音声を生成する．

本システムにおける STRAIGHT モーフィングを用いた特定話者音声生成は次の手順により行われる．

(1)　DB 内のモーフィングに用いる参照話者へ特徴点付与を事前に行う．
(2)　目標話者の声質を表すモーフィング率を推定する．
(3)　(2) で求めたモーフィング率に基づき，台詞 DB 内の参照話者に対してモーフィングを行い，目標話者の声質の台詞を生成する．

これらの手順のうち，最も重要な要素であるモーフィング率推定について，4.2.1 項およ

び 4.2.2 項において述べる．次に，4.2.3 項において，本システムで用いる多量の音声データに対して，ある程度自動的に特徴点を付与する手法について述べる．

4.2.1 制約なしモーフィング率推定

本論文では，複数話者の音声を用いた STRAIGHT モーフィングを拡張し，ある目標話者の声質になるようなモーフィング率の推定法を考える．本論文では，以下の式を満たすようにモーフィング率を決定する．

$$\begin{aligned}\hat{\boldsymbol{r}} &= \arg\min_{\boldsymbol{r}} |\boldsymbol{y} - \hat{\boldsymbol{x}}_{mrp}|^2 \\ &= \arg\min_{\boldsymbol{r}} \left| \boldsymbol{y} - \sum_{s=1}^{S} r_s \boldsymbol{x}_s \right|^2 \end{aligned} \tag{4}$$

$$\boldsymbol{r} = \left[r^{(1)}, r^{(2)}, \cdots, r^{(S)} \right]^{\top} \tag{5}$$

ここで，$\boldsymbol{y}$ は目標話者の特徴量ベクトル，$\hat{\boldsymbol{r}}$ は推定されたモーフィング率ベクトルである．また，$\top$ は転置を表す．この手法において，以下の式を最小とするように推定する．

$$\epsilon(\boldsymbol{r}) = \sum_{f,\tilde{f}=1}^{F} \sum_{\tilde{t}=1}^{\tilde{T}(\boldsymbol{r})} \left(y_{\tilde{t}}(f) - \boldsymbol{x}_{\tilde{t}}(\tilde{f})\boldsymbol{r} \right)^2 \tag{6}$$

ここで，$y_{\tilde{t}}(f)$ は $\tilde{t}$ 番目の時間特徴点上の目標話者の特徴量であり，F は標本化した際の最大周波数（ナイキスト周波数）を表す点である．また，$\boldsymbol{x}_{\tilde{t}}(\tilde{f}) = \left[x_{\tilde{t}}^{(1)}(\tilde{f}), x_{\tilde{t}}^{(2)}(\tilde{f}), \cdots, x_{\tilde{t}}^{(S)}(\tilde{f}) \right]$ はモーフィングに用いる S 人の参照話者の特徴量ベクトルである．ただし，$\tilde{f}$ は特徴点とモーフィング率 $\hat{\boldsymbol{r}}$ に基づいて伸縮した周波数成分である．また，$\tilde{T}(\boldsymbol{r})$ は時間伸縮後の発話区間フレーム長であり，以下の式により表される．

$$\tilde{T}(\boldsymbol{r}) = \sum_{s=1}^{S} r_s T_s \tag{7}$$

ただし，T_s は s 番目の参照話者の発話区間フレーム長である．

STRAIGHT に基づく音声モーフィングでは，音声特徴量，およびモーフィング後の時間周波数平面を同一のモーフィング率で制御するため，式 (6) における勾配は非線形であり，解析的に解くことが困難である．そこで，次式によりモーフィング率を逐次的に求める．

$$\hat{\boldsymbol{r}}_{n+1} = \hat{\boldsymbol{r}}_n - \alpha E_n \tag{8}$$

$$E_n = \left(\frac{\partial^2 \epsilon(\hat{\boldsymbol{r}}_n)}{\partial \hat{\boldsymbol{r}}_n^2}\right)^{-1} \frac{\partial \epsilon(\hat{\boldsymbol{r}}_n)}{\partial \hat{\boldsymbol{r}}_n}$$
$$= \left(\tilde{\boldsymbol{X}}_n^\top \tilde{\boldsymbol{X}}_n\right)^{-1} \tilde{\boldsymbol{X}}_n^\top \left(\tilde{\boldsymbol{Y}}_n - \tilde{\boldsymbol{X}}_n \hat{\boldsymbol{r}}_n\right) \tag{9}$$

ここで，$\epsilon(\hat{\boldsymbol{r}})$ は時間成分および周波数成分にかかるモーフィング率 $\hat{\boldsymbol{r}}_n$ を定数として，式 (6) と同様の計算で得られる誤差である．また，$\tilde{\boldsymbol{X}} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_{\tilde{T}(\hat{\boldsymbol{r}}_n)}^\top\right]^\top$ は n 回目の更新で得られたモーフィング率 $\hat{\boldsymbol{r}}_n$ により時間周波数伸縮を行った時間方向の特徴点情報を持つ特徴量であり，$\tilde{\boldsymbol{Y}} = \left[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \cdots, \boldsymbol{y}_{\tilde{T}(\hat{\boldsymbol{r}}_n)}^\top\right]^\top$ は時間方向の特徴点情報を持つ特徴量である．ただし，

$$\boldsymbol{X}_{\tilde{t}_n} = \left[\boldsymbol{x}_{\tilde{t}_n}^{(1)}, \boldsymbol{x}_{\tilde{t}_n}^{(2)}, \cdots, \boldsymbol{x}_{\tilde{t}_n}^{(S)}\right] \tag{10}$$
$$\boldsymbol{x}_{\tilde{t}_n}^{(s)} = \left[x_{\tilde{t}_n}^{(s)}(1), x_{\tilde{t}_n}^{(s)}(2), \cdots, x_{\tilde{t}_n}^{(s)}(\tilde{f}_n), \cdots, x_{\tilde{t}_n}^{(1)}(F)\right]^\top \tag{11}$$
$$\boldsymbol{y}_{\tilde{t}_n} = \left[y_{\tilde{t}_n}(1), y_{\tilde{t}_n}(2), \cdots, y_{\tilde{t}_n}(\tilde{f}_n), \cdots, y_{\tilde{t}_n}(F)\right]^\top \tag{12}$$

である．なお，本システムでは STRAIGHT スペクトルのみを用いてモーフィング率を推定し，得られたモーフィング率をほかの特徴量のモーフィングの際に適用している．

4.2.2　制約ありモーフィング率推定

予備検討においてモーフィング率の合計を 1 以上および 1 未満に設定したときに生成されるモーフィング音声が劣化することが確認されている．また，文献 24) において，モーフィング率を補外に設定した場合，音質が劣化することが報告されている．したがって，システムを運営するうえで STRAIGHT 音声モーフィングではモーフィング率の和が 1 でない場合，著しく音質が劣化する可能性がある．そのため，制約なしによるモーフィング率推定では音質が保証されないため，本システムに適用するのは難しい．そこで，以下のような制約をつけて，式 (4) を解く．

$$\boldsymbol{b}^\top \boldsymbol{r} = 1 \tag{13}$$
$$r_s \geq 0 \tag{14}$$

ここで $\boldsymbol{b} = [1, 1, \cdots, 1]^\top$ である．式 (4) および式 (13) を満たすため，Equality Constrained Least Squares（ECLS）[25] および Non-Negative Least Squares（NNLS）[26] を導入し，以下のように更新する．

$$\hat{\boldsymbol{r}}_{n+1} = \mathcal{L}_+ \left[\boldsymbol{Q} \left(\begin{bmatrix} \bar{r}_{n+1}^{(1)} \\ \bar{\boldsymbol{E}}_n \end{bmatrix} \bar{\boldsymbol{r}}_n - \alpha \begin{bmatrix} 0 \\ \bar{\boldsymbol{E}}_n \end{bmatrix} \right) \right] \tag{15}$$

$$\bar{E}_n = \left(\bar{\boldsymbol{X}}_n^\top \bar{\boldsymbol{X}}_n \right)^{-1} \bar{\boldsymbol{X}}_n^\top \left(\tilde{\boldsymbol{Y}}_{n+1} - \bar{\boldsymbol{x}}^{(1)} \bar{r}_{n+1}^{(1)} \bar{\boldsymbol{X}}_n - \bar{\boldsymbol{r}}_n \right) \tag{16}$$

ただし，

$$\mathcal{L}_+ [a] = \begin{cases} a, & \text{if } a \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

$$\boldsymbol{b} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix} \tag{18}$$

$$\boldsymbol{Q}^\top \boldsymbol{r} = \begin{bmatrix} \bar{r}^{(1)} \\ \bar{\boldsymbol{r}} \end{bmatrix} \tag{19}$$

$$\tilde{\boldsymbol{X}} \boldsymbol{Q} = \left[\bar{\boldsymbol{x}}^{(1)}, \bar{\boldsymbol{X}} \right] \tag{20}$$

$$\bar{r}^{(1)} = \boldsymbol{R}_{1\times 1}^{-1} \tag{21}$$

である．ここで，$\boldsymbol{Q}$ および $\boldsymbol{R}$ は $\boldsymbol{b}$ を QR 分解することで得られる直交行列，上三角行列であり，$\bar{r}^{(1)}$，$\bar{\boldsymbol{x}}^{(1)}$ は行列 $\boldsymbol{Q}$ をかけることで得られる 1 番目の参照話者のモーフィング率および特徴量，$\bar{r}$，$\bar{\boldsymbol{X}}$ は行列 $\boldsymbol{Q}$ をかけることで得られる 2 番目から S 番目の参照話者のモーフィング率および特徴量である．また，更新されたモーフィング率 $\hat{\boldsymbol{r}}_n$ は NNLS により負値となった場合，その値を 0 に置き換えて，和が 1 になるように正規化している．

4.2.3 動的計画法による特徴点付与

本システムでは多数の台詞発話を生成するために，複数の話者と大量の台詞発話で構成された DB を用いる．STRAIGHT 音声モーフィングでは新たな声質を生成するために音声特徴量に特徴点を付与する必要があり，限られた期間内ですべての特徴点を手動で付与することは難しい．そこで，これを解決するため，本システムでは動的計画法を用いての特徴点付与を次の手順により行った．

(1) あらかじめ DB 内から任意に選んだ基準となる話者の特徴量に対して手動による特徴点付与を行う．

(2) 自動特徴点付与を行う対象の話者の特徴量と基準話者の特徴量の間で DP マッチングを行い，対象話者に対して時間方向の特徴点の情報を付与する．

(3) 対象話者の時間方向の特徴点を付与されたフレームと，その特徴点に対応する基準話

者の時間方向の特徴点を付与されたフレームとの間で DP マッチングを行い，対象話者に対して周波数方向の特徴点情報を付与する．

5. 評価実験

5.1 音質に関する評価

システムの生成する台詞音声の音質を評価するために，主観評価実験を行った．被験者は，正常な聴力を有する成人男女 40 名（男女各 20 名）である．6 名の目標話者（男女 3 名ずつ）に対する出力結果を，1 つの音声あたり 20 名の被験者が評価するように割り当てた．被験者は，目標話者の音声と各手法により得られた音声とを比較し，5 段階の劣化評定尺度（DMOS）で評定値を付与する（0：劣化がまったく認められない，1：劣化が認められるが気にならない，2：劣化がわずかに気になる，3：劣化が気になる，4：劣化が非常に気になる）．なお，本評価実験における検知限の評定値を 0.5，許容限の評定値を 1.5 として評価を行う．つまり，評定値が 0.5 以下であれば検知限内，1.5 以下であれば許容限内であると判断する．提示した音声はサンプリング周波数 48 kHz，量子化ビット数 16 bit であり，発話内容は以下の 3 種類である．

SENT　ATR 音声データベース B セットの 503 文中ランダムに選択した 16 文の音声

FCSwoBGM　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 台詞から切り出した 39 音声サンプル

FCS　前述の発話内容 FCSwoBGM に，劇中の BGM/SE を重畳したもの

各手法に使用した話者データベースは 2.1 節に示したものを使用した．なお，音声モーフィング手法を用いた発話内容 FCS の音声を生成する際には，BGM/SE 重畳前の音声に対して手法を適用しモーフィング音声を生成したのち，BGM/SE を重畳した．

目標話者の音声に対する，目標話者の音声の STRAIGHT 分析合成音（ANASYN），類似話者選択結果音声（SELECT），モーフィング結果音声（MORPH）の平均評定値，および 95%信頼区間を**図 5** に示す．STRAIGHT 分析合成音については，BGM/SE のない発話内容 SENT，FCSwoBGM では許容限内の音質を確保し，発話内容 FCS では検知限内の音質を確保している．これにより，STRAIGHT 分析合成は本システムへの応用として十分な品質を有していることが確認できる．類似話者選択結果音声は，音質劣化の要因となるような音声加工を施していないため，原音声と比べても，検知限内の音質を確保できている．

目標話者の音声に対するモーフィング結果音声（MORPH）の評定値については，より詳細に分析するため，台詞ごとの平均評定値，および 95%信頼区間を**図 6** に示す．モーフィ

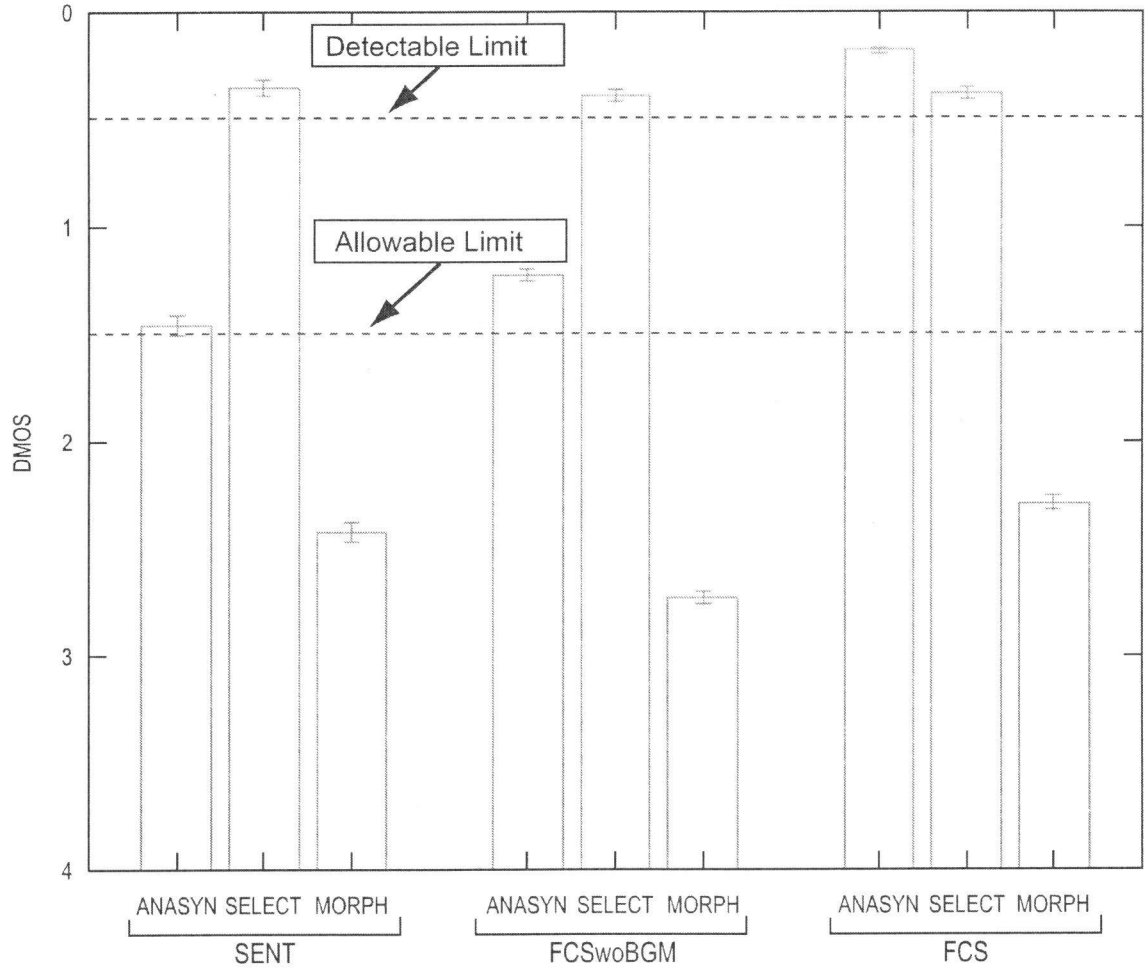


図 5 音質の主観評価実験結果

Fig. 5 Subjective evaluation result of voice quality.

ング音声の品質は，発話によってバラツキがあり，許容限内の音声も存在する．以上より，現状のモーフィング音声の品質では，上映前の音質検査が必須であることが確認された．

5.2 話者性に関する評価

5.2.1 客観評価実験

本論文で提案する STRAIGHT モーフィングを用いた特定話者音声生成手法の性能を評価するため，ケプストラム距離による客観評価を行った．本評価実験では，男性話者 13 名，および女性話者 17 名，各 2 種類の合計 60 種類の声質で発声された同一発話内容の音声を用いる．60 種類の音声のうち，1 つを目標話者の音声とし，残りの音声から，同一話者を除く目標話者の音声とのスペクトル距離が小さい音声を 2〜32 種類選び，これらをモーフィングに用いる参照話者の音声とする．目標話者の音声としては 60 種類すべて用いる．

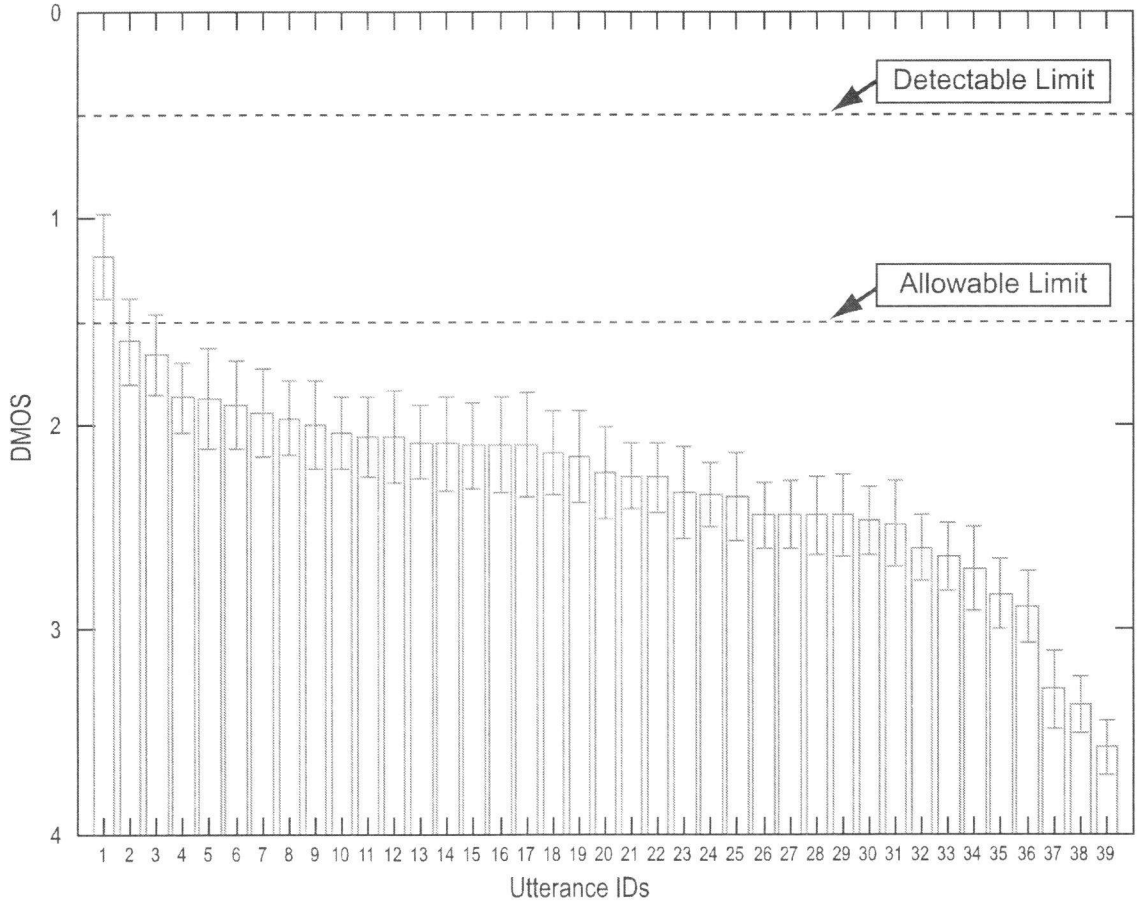


図 6　発話ごとの音質の主観評価実験結果（モーフィング音声）
Fig. 6　Subjective evaluation result of voice quality by utterances (speech morphing).

モーフィング率推定に用いた文は 1 文である．評価には同一の 1 文を用いる．モーフィングする際に必要となる特徴点はすべて手動で付与し，モーフィング率は 4.2.2 項に示した制約あり推定を用いて求める．また，モーフィング率の更新におけるステップサイズは $\alpha = 1$ とし，更新回数は 20 回として行う．

図 7 に客観評価実験結果を示す．図中の “Nearest speaker” は目標話者特徴量と距離が小さい参照話者特徴量とのケプストラム距離を表し，“Morphed speaker” は目標話者特徴量とモーフィング特徴量とのケプストラム距離を表す．図からモーフィング音声は話者選択を行う従来手法に比べ，目標話者とのケプストラム距離が近いことが分かる．また，モーフィングの際の参照話者人数を増やすことによりさらに近くなることが分かる．これにより，話者選択の場合と比較して，声質が目標話者により近いと考えられる．

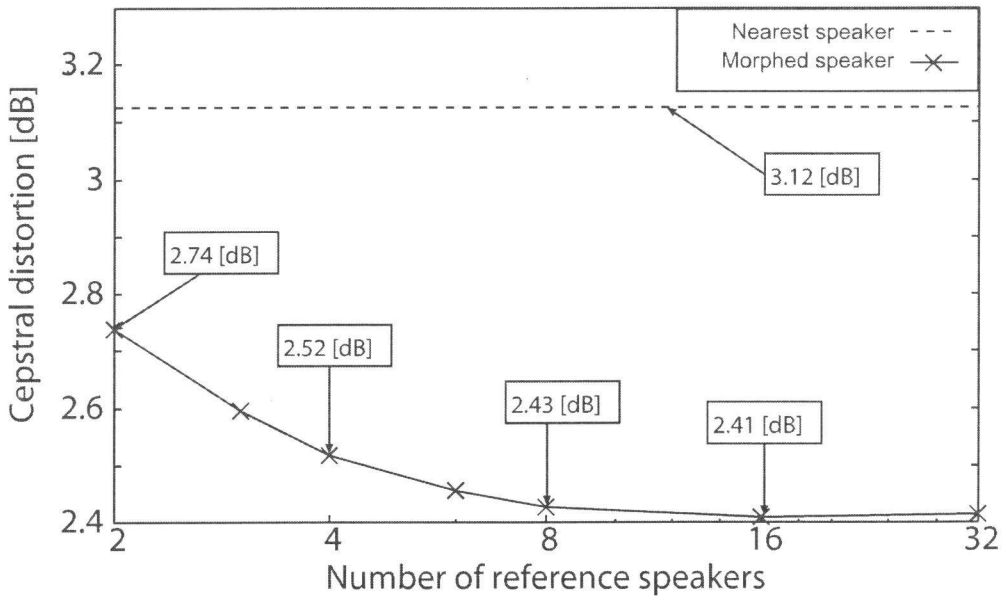


図 7 客観評価実験結果
Fig. 7 Result of objective evaluation.

5.2.2 主観評価実験

システムの生成する 2 手法（類似話者選択技術と音声モーフィング技術）の台詞音声の話者性を評価するために，主観評価実験を行った．被験者は，正常な聴力を有する成人男女 40 名（男女各 20 名）である．被験者は，目標話者の音声と 2 手法から得られる音声を提示し，どちらの手法の音声が目標話者の音声に似ているかを評価した．6 名の目標話者（男女 3 名ずつ）に対する出力結果を，2 グループ各 20 名の被験者が評価するように割り当て，一方のグループと他方のグループとで 2 手法から得られる音声の提示順序を逆にすることでカウンタバランスをとった．提示した音声はサンプリング周波数 48 kHz，量子化ビット数 16 bit であり，発話内容は以下の 2 種類である．

SENT ATR 音声データベース B セットの 503 文中ランダムに選択した 16 文の音声

FCS 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 台詞から切り出した 39 音声サンプルに，劇中の BGM/SE を重畳したもの

各手法に使用した話者データベースは 2.1 節に示したものを使用した．なお，音声モーフィング手法を用いた発話内容 FCS の音声を生成する際には，BGM/SE 重畳前の音声に対して手法を適用しモーフィング音声を生成したのち，BGM/SE を重畳した．

類似話者選択結果音声（SELECT）とモーフィング結果音声（MORPH）との話者性を比較した Preference Score，および 95%信頼区間を**図 8** に示す．図 8 より，モーフィング

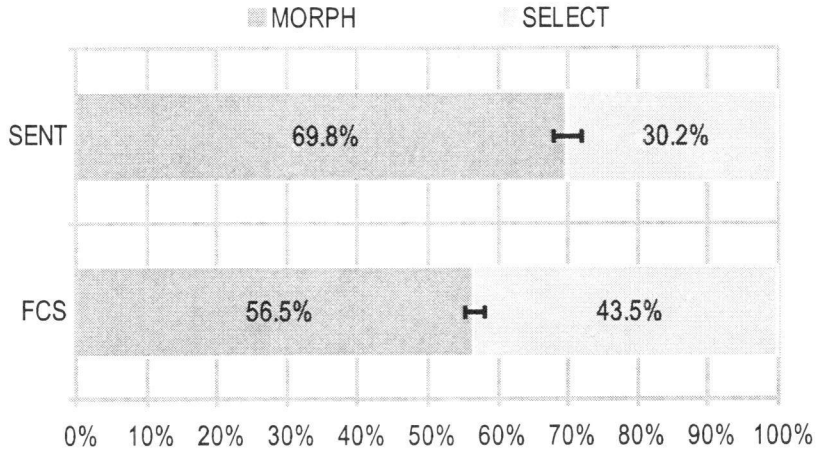


図 8　話者性の主観評価実験結果
Fig. 8　Subjective evaluation result of speaker similarity.

結果音声の方が明らかに目標話者に似ており，結果統合の際にモーフィング音声結果を優先することにより，システムの出力する台詞音声の話者類似性を改善できることを確認した．

以上の結果より，提案システムにおける結果統合の優先順位として，1 位を音声モーフィングに基づく手法，2 位を話者類似性評価に基づく手法とした．

5.3　システム統合に関する評価

日本未来科学館において，2009 年 3 月 20 日から 22 日まで，本システムを導入した FCS を展示した．主な体験者は親子づれの参加者で，3 日間で延べ 290 名以上に対する音声を生成した．体験は最大 10 名が同時に参加することができるように設計し，音声収録ノート PC 5 台，音声分析サーバ 3 台，類似音声選択サーバ 3 台，音声モーフィングサーバ 6 台，ファイルサーバ 1 台を準備した．音声モーフィングサーバのうち 1 台の PC はキュー管理サーバとして兼用した．処理時間算出に関連のある音声分析サーバ，類似音声選択サーバ，音声モーフィングサーバについては，CPU：AMD Phenom X4 9950（Quad-Core，2.6 GHz），メモリ：4 GB のものを使用した．なお，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” の登場キャラクタは 20 名（110 音声サンプル）であるため，本システムでは半分のキャラクタ 10 名（61 音声サンプル）を処理対象とし，残り半分のキャラクタ 10 名（49 音声サンプル）については体験者の音声が与えられないものとして標準音声を用いた．音声モーフィング処理については，処理時間・処理 PC 台数の観点から，対象キャラクタを 4 名（17 音声サンプル），モーフィングに用いる参照話者の人数を 8 名とし，モーフィング率推定処理とモーフィング音声生成処理は並列に実行した．体験者の音声収録は，極力雑音の混入を避けるため，簡易の収録室内でヘッドセットマイクを用いた．

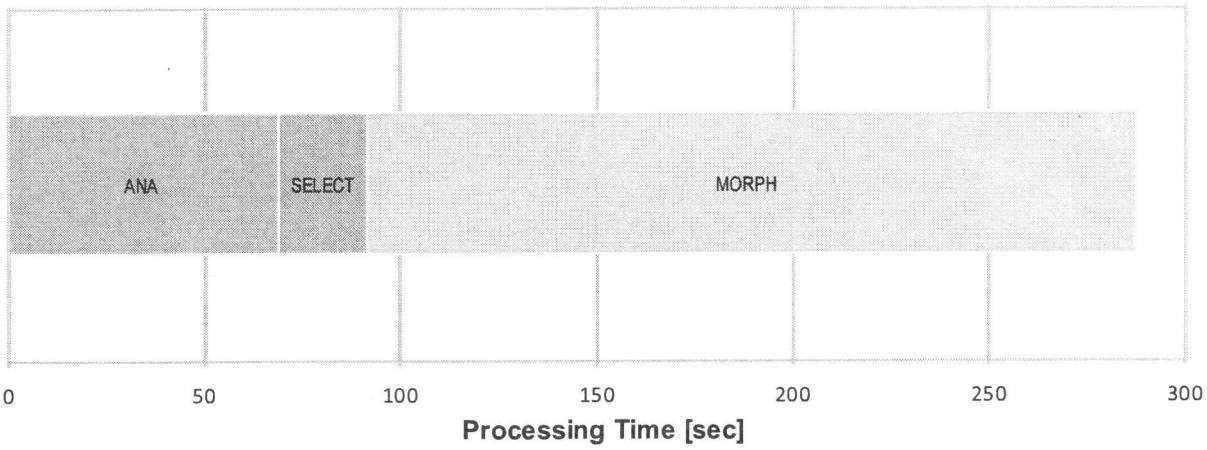


図 9 処理時間
Fig. 9 Processing time.

音声収録に際しては，音声モーフィング対象キャラクタを割り当てられた体験者から優先的に収録した．つまり，体験者を同時音声収録可能人数（音声収録ノート PC 台数）である 5 名ずつ，2 グループに分けて収録を行い，最初の 5 名の収録に音声モーフィング対象キャラクタを割り当てられた 4 名の体験者が含まれるようにした．

2009 年 3 月 22 日の 10 回の上映（参加者延べ 100 名）について，体験者から収録した音声（平均収録時間 273 秒，平均音声継続長 2.99 秒）の音声分析開始から各処理（ANA：音声分析，SELECT：類似音声選択，MORPH：音声モーフィング）の終了までの経過時間を計測し，平均処理時間を算出した結果を**図 9** に示す．各処理終了までの平均時間が音声分析終了 69.14 秒，話者選択終了 91.61 秒，モーフィング終了 287.55 秒であった．また，2.5 節に示した結果統合処理は上映 2 分前の時点で実行し，その処理時間は 1 名の聴取者によるモーフィング音声の音質評価時間も含め，2 分以内であった．以上より，FCS の準備時間の条件 15 分以内に処理できていることが確認できる．同時体験者数や音声モーフィング処理対象キャラクタ数の増加については，音声収録や処理に用いる PC の増加，および聴取による音質確認を行う人員の増加により，15 分以内に処理できると考えられる．

結果統合の評価として，2009 年 3 月 22 日の 10 回の上映について，モーフィング音声，類似話者選択音声，標準音声がそれぞれどれくらい選択されたかを調査した．その結果，標準音声は平均 49.0 音声サンプル，類似話者選択音声は平均 57.7 音声サンプル，モーフィング音声は平均 3.3 音声サンプルであった．類似話者選択対象であった 61 音声サンプルについては，時間内に処理が完了しており，標準音声が選択されることはなかった．また，類似話者選択対象であり，かつ音声モーフィング処理対象であった 17 音声サンプルのうち，音

声モーフィングが採用されたのは平均 3.3 音声サンプルであることから，音声収録ノート PC，および各種サーバを増加し，すべての台詞に対して音声モーフィングを適用した場合は，足立らの従来システム[2] と比べ 19.4%（3.3/17）話者類似性を改善できると考えられる．本手法がなければ，それら 17 音声サンプルはすべて類似話者選択音声が選択されるが，それよりも話者性の観点から良いものが選ばれたのだから，システム全体として品質向上の可能性があるといえる．さらに，類似話者選択対象であり，かつ音声モーフィング処理対象であった 17 音声サンプルのうち，音声モーフィングが採用されなかったもののほとんどは，音質確認の時点で劣化が認められるものが破棄されたためであり，音声モーフィングの音質改善により，システム全体の話者性改善が期待できる．

6. 議　　論

音声モーフィングでは，特徴点がモーフィング時の対応点となるため，特徴点付与の精度が音質に大きく影響する．今回は，多数の台詞音声の効率的な整備を行うため，動的計画法を利用して，基準話者以外の台詞音声に対する特徴点を自動付与した．しかし，基準話者とその他の話者との個人差などから，局所的な特徴点のずれを生じさせる可能性がある．モーフィング率と音質の関係については，文献 24) において，補間であれば十分な音質が得られることが報告されている．したがって，本論文の制約付き重み推定を用いることで音質の劣化は抑えられると考えられる．一方，特徴点付与においては，モーフィングに用いるすべての参照話者の音声に対して特徴点の持つフォルマント情報や音素情報を一致させる必要がある．したがって，特徴点情報が一致していない参照話者を用いてモーフィングを行う場合，音素環境やフォルマント位置の不一致によりモーフィング音声の音質は劣化すると考えられる．また，特徴点付与は重み推定にも影響を与えると考えられる．つまり，特徴点情報が一致していない参照話者を用いて重み推定を行う場合，得られる重みは目標話者の声質を正しく反映できないと考えられる．以上の点から，特徴点付与の精度は音質により大きく影響すると考えられる．理想的には，システム構築にかけられる時間・コストに応じて，今回の特徴点自動付与結果をベースに手作業で特徴点を調整することが望ましい．手作業による特徴点付与作業は非常に時間と手間のかかる作業であるため，特徴点付与作業を支援する技術・環境の整備や，特徴点自動付与技術の高精度化が今後の課題としてあげられる．

話者により文末表現の「です」「ます」における母音/u/の無声化がみられるように，一般には同じ文章を発話しても話者により音素ごとの有声・無声のパターンは異なる．また，文中のポーズの位置も異なる．現状の音声モーフィング手法では，同一文章を発話した音声

では，文中のポーズの位置や有声・無声のパターンが一致していることを仮定した手法となっている．ポーズの位置や有声・無声のパターンの不一致が，特徴点付与における精度低下の要因やモーフィング音声の音質劣化の要因となりうるため，音声モーフィングに使用する話者データベースはこれらの情報が一致するように整備しなければならない．一方で，効率的に音声モーフィングのための話者データベースを整備するには，ポーズの位置や有声・無声のパターンの一致といった収録上の制約は少ない方が望ましい．音声モーフィング技術に関するこれらの制約を緩和することが，今後の課題としてあげられる．

類似話者選択技術における結合係数の最適化は，現状のシステムでは成人女性 36 名のデータに付与された順位をもとに行われている．一方，本システムの利用者としては，成人男女や子供の利用も想定されるが，成人女性のデータで最適化した結合係数が成人男性や子供においても有効であるかどうかは実験されていない．また，被験者の負担の制限から，評価対象話者数や発話文章数，順位付与被験者数など小規模な実験による技術評価にとどまっている．しかし足立らの報告にあるように，本技術を導入したシステムは，応用事例としては有効であることが示されている[2]．今後は，性別や年齢に関する依存性や文章依存性などの観点から，より詳細に結合係数最適化の頑健性やシステムとしての影響に関する調査を進めることが課題である．

話者データベース以外はコンテンツに依存していないため，話者データベースをコンテンツに合わせて差し替え，データを整備することで，他のコンテンツに応用することも可能である．同時体験人数についても，処理対象となる音声は増加するが，同様の枠組みで対応できる．さらに，複数来場者の音声は並列に処理可能であるため，処理用 PC を増加することで，準備時間を維持しつつより体験人数の多い作品への対応も可能である．

現状のシステム運用においては，人手による音声モーフィング出力の音質検査によって音質の保障を行っている．本システムで対象とした音声の音質評価時間は，統合処理を含めても 2 分以下であり，少ない作業コストでシステムが出力する台詞音声の話者類似性が改善できる利点は大きいと考えられる．しかし，システムの拡張にともない，音質検査対象の手法や音声が増えるにともない検査作業時間も増加する．複数人数で検査作業を分担することもできるが，将来的には作業効率化のため音質評価も自動化されることが望ましい．

7. ま と め

視聴者から収録した少量の音声を用いて，視聴者の声に似た台詞音声を生成するため複数手法を統合し，生成された台詞音声をシーンに合わせて同期再生することで，視聴者の声

の特徴をキャラクタに反映させるシステムを提案した．提案システムは，エンターテインメントシステムでの使用を想定し，映像作品に期待される音質を確保し，視聴者の収録などの負担を極力軽減し，短い待ち時間で，映像に合わせて台詞音声を同期出力する，頑健性・柔軟性を有するシステムを設計した．FCS において，話者データベースから視聴者と声が似た話者を選択する手法（類似話者選択技術）と，複数話者音声を混合することで視聴者の声に似た音声を生成する手法（音声モーフィング技術）を組み合わせたシステムを開発し，複数処理を並列化することで，上映準備時間の要求条件を満たした．実環境を想定して BGM/SE を重畳した音声で，従来手法である類似話者選択技術より得られる音声と，提案法で導入した音声モーフィング技術より得られる音声を主観評価実験により評価した結果，Preference Score で 56.5%のモーフィング音声が目標話者の音声に似ていると判断され，音声モーフィングを組み合わせることでシステムが出力する台詞音声の話者類似性を改善できることを示した．

謝辞　本研究は文部科学省の科学技術振興調整費「新映像技術ダイブイントゥザムービーの研究」の助成を受けた．

参 考 文 献

1) Maejima, A., Wemler, S., Machida, T., Takebayashi, M. and Morishima, S.: Instant Casting Movie Theater: The Future Cast System, *IEICE Trans. Inf. Syst.*, Vol.E91-D, No.4, pp.1135–1148 (2008).
2) 足立吉広，大谷大和，川本真一，四倉達夫，森島繁生，中村　哲：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，情報処理学会論文誌，Vol.49, No.12, pp.3908–3917 (2008).
3) 河井　恒，戸田智基，山岸順一，平井俊男，倪　晋富，西澤信行，津崎　実，徳田恵一：大規模コーパスを用いた音声合成システム XIMERA，電子情報通信学会論文誌，Vol.J89-D, No.12, pp.2688–2698 (2006).
4) 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正：HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，電子情報通信学会論文誌，Vol.J83-D-II, No.11, pp.2099–2107 (2000).
5) Toda, T., Black, A. and Tokuda, K.: Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.8, pp.2222–2235 (2007).
6) Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure

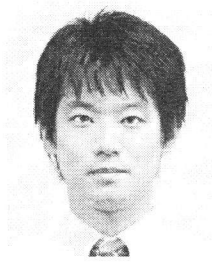
in sounds, *Speech Commun.*, Vol.27, No.3-4, pp.187–207 (1999).
7) 伊藤　玄，葦苅　豊，實廣貴敏，中村　哲：音声認識統合環境 ATRASR の概要と評価報告，日本音響学会秋期研究発表会講演論文集，No.1-P-30 (2004).
8) 松田繁樹，實廣貴敏，中村　哲，石井カルロス寿憲，神田崇行：コミュニケーションロボットにおける音声認識性能の評価，日本音響学会秋期研究発表会講演論文集，No.2-P-22 (2005).
9) 四倉達夫，川本真一，松田繁樹，中村　哲：iFACe：デジタルアニメ声優体験システム，情報処理学会論文誌，Vol.49, No.12, pp.3847–3858 (2008).
10) Kawahara, H. and Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '03*), Vol.1, pp.256–259 (2003).
11) Slaney, M., Covell, M. and Lassiter, B.: Automatic audio morphing, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '96*), pp.1001–1004 (1996).
12) Adachi, Y., Kawamoto, S., Morishima, S. and Nakamura, S.: Perceptual similarity measurement of speech by combination of acoustic features, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '08*), pp.4861–4864 (2008).
13) Reynolds, D. and Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech and Audio Processing*, Vol.3, No.1, pp.72–83 (1995).
14) 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，日本音響学会春季講演論文集，pp.443–444 (2007).
15) 永嶋育美，高際光晴，齋藤　裕，柳川博文，長尾優次：人の話者識別における音声の類似性の検討，日本音響学会春季講演論文集，pp.737–738 (2003).
16) 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，日本音響学会春季講演論文集，pp.441–442 (2007).
17) 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，日本音響学会秋季講演論文集，pp.263–264 (1996).
18) 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，日本音響学会春季講演論文集，pp.363–364 (2000).
19) 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).
20) Atal, B.S. and Hanauer, S.L.: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, *The Journal of the Acoustical Society of America*, Vol.50, No.2B, pp.637–655 (1971).
21) 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単語認識，

日本音響学会誌，Vol.27, No.9, pp.483–490 (1971).
22) 齋藤堯幸，宿久　洋：関連性データの解析法—多次元尺度構成法とクラスター分析法，共立出版 (2006).
23) 高橋　徹，西　雅史，入野俊夫，河原英紀：多重音声モーフィングに基づく平均声合成の検討，日本音響学会春季講演論文集，pp.229–230 (2006).
24) 松井九美，河原英紀：STRAIGHT による感情モーフィング音声の知覚と変換関数の構造について，電子情報通信学会技術研究報告，No.SP2003-27, pp.63–68 (2003).
25) Golub, G.H. and Van Loan, C.F.: *Matrix Computations* (*Johns Hopkins Studies in Mathematical Sciences*), chapter 12, pp.580–587, The Johns Hopkins University Press (1996).
26) Bro, R. and Jong, S.D.: A fast non-negativity-constrained least squares algorithm, *Journal of Chemometrics*, Vol.11, No.5, pp.393–401 (1997).

(平成 21 年 4 月 20 日受付)
(平成 21 年 11 月 6 日採録)

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 年同大学院博士後期課程修了．博士（情報科学）．2005 年より（株）国際電気通信基礎技術研究所（ATR）．2009 年より（独）情報通信研究機構（NICT）．現在，NICT 知識創成コミュニケーション研究センター専攻研究員，および，ATR メディア情報科学研究所客員研究員．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．



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工学専攻博士前期課程修了．現在，同大学院博士後期課程在学中．2006〜2009 年 ATR 音声言語コミュニケーション研究所研修研究員．主として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．

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．2003 年同大学院博士課程修了．博士（工学）．2003 年（株）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研究員．現在，（株）オー・エル・エム・デジタル研究開発部門ソフトウェアエンジニア，（株）国際電気通信基礎技術研究所メディア情報科学研究所客員研究員．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に関する研究開発に従事．電子情報通信学会学術奨励賞，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 年東京大学大学院工学系研究科電子工学専攻博士課程修了，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 年早稲田大学先進理工学部応用物理学科教授，現在に至る．1994 年から 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 2008 年国際電気通信基礎技術研究所客員研究員．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．NICT 客員研究員を併任．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 年シャープ（株）中央研究所，情報技術研究所勤務．1992 年京都大学博士（工学）．1994 年奈良先端大情報科学研究科助教授．1996 年米 Rutgers 大学 CAIP センター客員教授．2000 年（株）国際電気基礎技術研究所および（株）ATR 音声言語研究所第一研究室長．2001 年（株）国際電気基礎技術研究所音声言語コミュニケーション研究所第一研究室長．2005 年同執行役員所長，2006 年取締役，2008 年 ATR フェロー，2006 年（独）情報通信研究機構知識創成コミュニケーション研究センター音声言語 GL（兼務），2007 年同上席研究員（兼務）．現在，（独）情報通信研究機構知識創成コミュニケーション研究センター副研究センター長，MASTAR プロジェクトリーダ，音声コミュニケーション GL，および，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．

情報処理学会論文誌　Vol. 51　No. 2　250–264 (Feb. 2010)