情報処理学会論文誌 Vol. 49 No. 12 3847–3858 (Dec. 2008)

iFACe：デジタルアニメ声優体験システム

四倉達夫[†1] 川本真一[†1]
松田繁樹[†1] 中村哲[†1]

声優経験のない参加者が，プロの声優のアフレコした発話アニメーションに近いリップシンク精度で，デジタルアニメキャラクタの発話と同期したアニメーションを体験できる，インタラクティブ発話合成システム *iFACe* を提案する．iFACe はリップシンクアニメーションを素早く生成するため，参加者から収録した台詞音声のタイミングに合わせて CG キャラクタ発話アニメーションを生成する，プレスコ方式を用いている．本システムは，参加者が選択した台詞情報と台詞に対応した音声を入力とし，雑音環境下でも推定精度の高い音素アラインメントを行い音素と音素継続長を求める．次にさまざまなスタイルのカートゥーンキャラクタにあうキーフレームの作成を行う．出力したキーフレームから，ブレンドシェープアプローチによる GPU を使ったリアルタイム発話アニメーションと音声を同期し出力を行う．本システムは日本科学未来館に 5 日間のデモンストレーションを行い，主観評価実験から，74%の回答者が，ゲームとして声優体験システムで遊んでみたいと示し，エンタテインメントシステムとしての有効なコンテンツであることが示された．

iFACe: Interactive Facial speech Animation Control systEm for 3D Cartoon Characters

Tatsuo Yotsukura,[†1] Shin-ichi Kawamoto,[†1]
Shigeki Matsuda[†1] and Satoshi Nakamura[†1]

In this paper, we propose a novel interactive lip-sync animation system for entertainment that works with players' voices and transcriptions as input and provides following: Robust speech recognition for a wide range of consumers in noise environments; smoothing lip-sync animation for cartoon characters; and blend-shaped based technique common in CG production real-time lip-sync animation on graphics hardware. We demonstrated and evaluated our system at National Museum of Emerging Science and Innovation (Miraikan) for five days. The evaluated results showed that our system was effective contents for entertainment use.

1. はじめに

本稿では，デジタルアニメ映像に対して声優のスキルを必要とせず，発話者の声とキャラクタの発話アニメーションとの同期（リップシンク）を実現可能な，エンタテインメント向けインタラクティブ発話合成システム iFACe を提案する．アニメ映像における台詞音声は，キャラクタやシーンにあった演技に加え，キャラクタの口の動きと音声が同期していることが求められる．この要求を満たすための音声収録手段として，「プレスコ（音声を先に収録し，音声に合わせて映像を制作する）」と「アフレコ（映像に合わせて後から音声を吹き込む）」がアニメ制作現場で用いられている．Disney[1)] などハリウッドプロダクションでは発話者の音声からキャラクタの発話アニメーションを作成する「プレスコ」でリップシンクを実現する．プレスコは，発話の始端終端だけでなく，発話中の音素に合う口形状を制作するため，リップシンク精度の高いアニメーションを実現できる反面，膨大な制作コスト・時間が必要となる．またプレスコを採用することで，発話者がアニメーションとの同期を意識することなく音声を収録することができる．

iFACe はゲームセンタや職業体験ができるアミューズメント施設で，子供でも簡単に音声収録ができ，短時間に 3DCG キャラクタと参加者の声とのリップシンクを体感可能なシステムの実現を目標とする．またエンタテインメントとして，参加者から収録した音声に合わせて，さまざまな CG キャラクタの発話アニメーションが，参加者の家族や友人とともに鑑賞できる．そして参加者は，プロの声優が持つ “映像や音声に合わせて発話者の発話速度・タイミングを微調整するスキル” を必要とせず，参加者の演技のみに集中して音声を収録し，収録した音声に同期した発話アニメーションをその場で楽しむことができる．既存のコンシューマ用コンテンツにない新しいゲーム性を持つ．さらに本システムを基盤技術としたアバタや，エージェントシステムをはじめとしたさまざまなアプリケーションが構築できるよう，機能ごとにモジュール化されたシステム設計を有する．これらを実現するため，次に示す条件：1) 参加者が簡単かつ短時間に音声収録，2) アミューズメント施設のような雑音環境下，子供音声に対応したリップシンク精度の高いアニメーション，3) トポロジの異なるさまざまな CG キャラクタによるリアルタイム発話アニメーションを生成，4) さま

†1 ATR 音声言語コミュニケーション研究所
ATR Spoken Language Communication Research Laboratories

© 2008 Information Processing Society of Japan

ざまなアプリケーションへ利用可能なシステム設計を解決する必要がある．

iFACe は上記条件に対して以下の特徴を持つ．

プレスコ音声による声優体験：音声収録方法は「アフレコ」もしくは「プレスコ」があるが，iFACe は「プレスコ」を用いている．「アフレコ」は事前に，台詞に合うキャラクタ発話アニメーションを作成することで，簡単に動画を再生することができる．しかし，声優経験のない参加者は作成した動画に合わせ発話する必要があり，リップシンクされた音声を収録するためには膨大な時間を要する．一方「プレスコ」は，音声に合った CG キャラクタのリップシンクアニメーションを作成しなければならないが，参加者は発話タイミングを考えることなく，演技に集中して収録することができる．iFACe は，参加者の台詞音声を先に収録することで，収録の負担を軽減させる．さらに，その音声の発話タイミングを分析し，分析結果から短時間に CG キャラクタのリップシンクアニメーションを自動生成することで，条件 1) である参加者が簡単かつ短時間に音声収録できる「プレスコ」を採用した．提案システムに関連した開発物として，ナムコが開発したアフレコゲーム[2]，またタイトーが開発したアフレコ体験システム[3] があるが，いずれもリップシンクのズレをスコア化し，アフレコ体験の難しさをゲーム化したものである．iFACe はプレスコによる声優体験システムであり，参加者が収録に必要なスキルを必要とせずすぐに利用することができる．

アミューズメント施設利用を考慮した音声分析：台詞音声を入力とし，音声に対応した音素・音素継続長を出力する，音素アラインメントは Julius[4] をはじめさまざまなシステムで実現することができる．iFACe は 2) 雑音環境下，子供音声を入力としたリップシンク精度の高いアニメーションを実現するため，耐雑音性能に優れ，フレキシブルな音響モデルに対応した，当研究所で開発された音声認識ソフトウェアを用いて音素アラインメントを行った．マイクから入力された音声は，パーティクルフィルタを用いた背景雑音の除去手法により雑音抑圧される．また，子供用音響モデルをシステムに追加することで，大人用音響モデルだけでは認識しにくい子供の声にも対応した．

キャラクタのスタイルに合った発話アニメーション：分析結果である音素・音素継続長を抽象的な CG キャラクタへ反映すると，身体などの動きと比べ口の動きだけ素早く動き目立ってしまう，いわゆる “うるさい” 発話アニメーションになる場合がある．従来研究ではさまざまなリップシンクアニメーション技術[5)–8)] が提案されている．iFACe では，カートゥーンキャラクタのスタイルに合うリップシンクアニメーション技術[9] を用いることで，違和感のない発話アニメーションを提供することができる．またリップシンクに使用するキャラクタは無表情＋「あ」「い」「う」「え」「お」の 5 母音のみを用意すれば，さまざまな視覚素を自動作成することが可能となる．

GPU ベースリアルタイムアニメーション：条件 3) を満たす，参加者から収録した音声に同期するさまざまなキャラクタによるリップシンクアニメーションを短時間で生成・表示するため，iFACe は 3DCG アニメ制作に用いられるブレンドシェープ法を用いて，リップシンクアニメーションを生成した．無表情＋「あ」「い」「う」「え」「お」の母音に対応したジオメトリを GPU メモリ上へ展開し，キーフレームベースでのブレンドシェープ演算を GPU 上で処理することで，10 万ポリゴン以上の高精細ポリゴンで構成されたキャラクタを 60 fps でアニメーションすることができる．

シンプルなインタフェース：条件 1) を満たすため，短時間に音声を収録し子供たちが簡単に iFACe を操作できるように，極力複雑な操作を行わないように，タッチパネルディスプレイを用いて 8 つの台詞選択ボタンから，参加者が簡単に選択・収録できる GUI を設計した．

汎用性の高いシステム設計：条件 4) さまざまなアプリケーションへ利用可能なシステム設計を行うため，4 つの独立した機能モジュール（フロントエンドモジュール，音声分析モジュール，リップシンク用キーフレーム作成モジュール，キャラクタ生成モジュール）を互いに TCP/IP ソケット通信で接続してシステムを構築する．これにより，用途に応じて，必要なモジュールが簡単に利用でき，各モジュールを他システムへ利用する場合でも，簡単に利用することができるシステム拡張性を確保する．

関連研究として Galatea FSM をはじめとするエージェント対話システム[10]，アバタ対話システム[11] があげられる．Galatea FSM，アバタ対話システムにおける CG キャラクタ生成では，実際の顔写真をさまざまな表情・口形状が定義された標準顔ワイヤフレームモデルへマッピングを行い，制御コマンドにあわせてブレンドシェープによる表情変化・リップシンクアニメーションを行う．iFACe は一般的な CG ソフトウェアで作成された，ポリゴン形状の CG キャラクタモデルと，キャラクタ用口形状モデルを用意することで，さまざまなキャラクタを表示することができる．また関連研究は CPU でブレンドシェープの演算を行っているのに対し，iFACe は GPU を用いてバーテックスシェーダ上でブレンドシェープ演算を行うため，高精細ポリゴンで構成されたキャラクタをリアルタイムに制御することが可能となる．音声分析に関して，アバタ対話システムでは，ニューラルネットを用い 5 母音を推定することで，リアルタイムリップシンクアニメーションを実現している．iFACe の音声分析部分は音声認識システムを利用し，Viterbi アラインメントで音素，音素継続長の推定を行うことで，高精度なリップシンクアニメーションを構築することができる．

本論文の構成として，2 章では iFACe の概要について述べ，3 章では雑音環境下で動作

© 2008 Information Processing Society of Japan

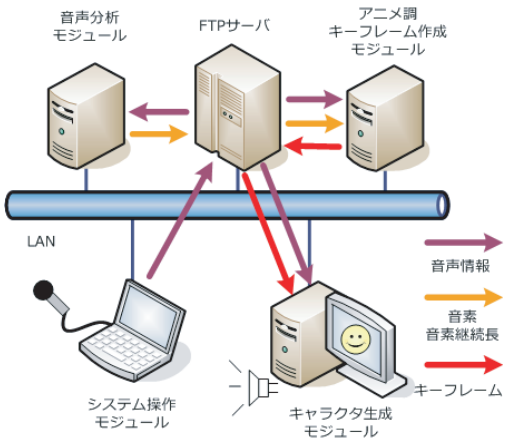

図 1　システム構成
Fig. 1　System overview.

図 2　iFACe 台詞選択ウインドウ
Fig. 2　iFACe dialog window for selection of transcription.

可能な柔軟な音声分析モジュールについて述べる．4 章ではアニメ調キーフレーム作成モジュールの構築について，5 章ではリアルタイムリップシンクアニメーションの生成モジュールについて紹介する．そして 6 章では，提案システムの評価実験，7 章では評価結果に対するディスカッション，8 章ではまとめについて述べる．

2. システム概要

iFACe のシステム構成を**図 1** に示す．システム操作，音声分析，アニメ調キーフレーム作成，キャラクタ生成の 4 モジュールで構成されている．システムフローは次のとおりである．

(1)　まずシステム操作用モジュールのインタフェースで，ユーザはタッチパネルで発話したい台詞を選択し，音声を入力する．**図 2** にラベル付き音声入力の台詞選択画面を示す．

(2)　入力された音声データは音声分析モジュールで処理を行い，音素アラインメント（音素，および音素継続長の推定）を行う．

(3)　分析後，アニメ調キーフレーム作成モジュールでキャラクタのスタイルに合うリップシンク用キーフレームへ変換し，キャラクタ生成モジュールへ送られる．

(4)　キャラクタ生成モジュールでは，あらかじめ GPU メモリへ読み込んだキャラクタのジオメトリと口形状データからキーフレームデータを用いて，リップシンクアニメーションを生成する．

データの受け渡しは FTP サーバを経由して行われ，各モジュールの分析の開始・終了・データ送受信命令・エラー制御などのモジュール間通信は，TCP/IP ソケット間通信により行われる．

3. 雑音環境下で動作可能な音声分析モジュール

実環境では，さまざまな種類の雑音が存在する．空調などの定常な雑音だけでなく，iFACe の利用を想定しているアミューズメント施設では，ゲームの効果音や大勢の話し声など，時々刻々と変化する非定常な雑音が多く含まれる．このような背景雑音は，音声分析の精度の劣化につながる．また，本システムは，大人だけなく子供の利用を想定している．話者の世代の違いは，声道長の違いによるスペクトルの伸縮や，発話スタイルの違いを含んでいる．このような話者の世代の違いに起因する発話スタイルの違いは，雑音と同様に音声分析の精度の劣化につながる．

本章では非定常な背景雑音や話者の世代の変化に頑健な音声分析モジュールについて述べる．音声分析モジュールは，あらかじめ与えられた発話テキストに対応する音素列を用いて，個々の音素の継続時間を推定する．音声分析モジュールの構造を**図 3** に示す．本モジュールは，大きく 2 つの技術から構成される．雑音抑圧部では，パーティクルフィルタによる非定常背景雑音の抑圧[12] が行われる．10 ms ごとに背景雑音と音声のスペクトラムが

© 2008 Information Processing Society of Japan

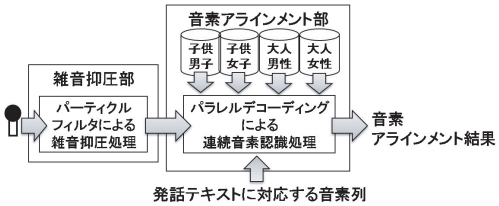

図 3　音声分析モジュールの構造
Fig. 3　Overview of speech analysis module.

逐次推定され，それらを用いたウィーナフィルタが計算される．入力された雑音重畳音声に対して，計算されたウィーナフィルタを適用することで背景雑音が抑圧される．

音素アラインメント部では，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[13] が行われる．発話テキストに対応する音素列の制約を基に，大人用の音響モデルと子供用の音響モデル，各々を用いて音素アラインメントが行われる．得られた音素アラインメント結果の音響スコアを比較し，最も高いスコアの得られた結果が出力される．このような選択を行うことにより，大人と子供の声の両方に対して頑健な音素アラインメントを行うことが可能となる．

3.1 実験条件

音声分析性能評価のため，大人音声評価用データ，20 歳代から 40 歳台の 12 名（男性 6 名，女性 6 名）各話者 5 文章（合計 60 文章）および，子供音声評価用データ，小学校 4 年生から 6 年生の男女（男子 3 名，女子 3 名），各話者 5 文章（合計 30 文章）の音声データを準備した．発話文章のタスクは旅行会話（BTEC）である[14]．大人用音響モデルは，ATR 旅行対話データベース TRA および，スケジューリング会話データベース APP，音素バランスの読み上げ音声 TRA-BLA と APP-BLA を用いた[14]．総発話文章数は，約 17 万である．子供用音響モデル学習データとして，名古屋大学において収録された CIAIR-VCV（子供の声データベース）[15] を用いた．音響特徴量は，12 次元 MFCC および，12 次元 ΔMFCC，対数パワーの計 25 次元である．音素列は発話テキストから生成した．

評価実験では，音素アラインメント中の個々の母音の開始時刻に対して，検知限と呼ばれる，人が不自然に感じる映像と音声のズレ幅（音進み 45 ms，音遅れ 125 ms）を評価尺度として用いた[16]．検知限を超えるズレに対して誤り，検知限に収まるズレに対して正解として，個々の母音のアラインメントに対する正解アラインメント率を計算した．正解とする個々の音素の開始時刻は，クリーン音声に対する Viterbi アラインメントにより求めた．

表 1　雑音抑圧の有無に対する評価結果
Table 1　Performance of the system with/without noise suppression.

	騒音レベル（dBA）				
	63	68	73	78	83
抑圧なし（%）	99.16	98.99	97.32	96.31	90.94
抑圧あり（%）	99.16	99.16	98.99	98.66	96.81

アラインメントに用いる音響モデルとしては，環境依存モデル，tied-state HMM を採用した．文献 17) の報告によると，HMM によるアラインメント結果の誤差平均は人が不自然に感じる映像と音声のズレ幅（音進み 45 ms，音遅れ 125 ms）に収まるため，ノイズ環境のないクリーンな音声を入力としたアラインメント結果は，リップシンクアニメーションを生成するために十分な精度を有する．

3.2 背景雑音に対する頑健性評価実験

背景雑音に対する頑健性を評価するため，アミューズメント施設で収録した雑音を用いた．騒音レベルは 83 dBA であった．ゲームの効果音や音楽，人の話し声などが含まれている．収録した雑音を用い，計算機上で重畳することにより，83 dBA，78 dBA，73 dBA，68 dBA，63 dBA の環境下で発話された雑音重畳音声を準備した．パーティクルフィルタによる雑音抑圧および，大人用音響モデルのみを持つモジュールを構築した．比較実験として，雑音抑圧を行わない場合の評価を行った．実験結果を**表 1** に示す．

表に示すように，雑音抑圧を行うことにより，雑音抑圧を行わない場合と比較して高い正解アラインメント率が得られた．83 dBA といった高騒音下で雑音抑圧を行わない場合 90.94%であるのに対し，雑音抑圧処理の導入により 96.81%まで正解アラインメント率が改善した．

3.3 話者の世代に対する頑健性評価実験

次に，話者の世代に対する頑健性を評価するため，大人音声と子供音声の両方を用いて実験を行った．パーティクルフィルタによる雑音抑圧および，大人用音響モデル，子供用音響モデルの両方を用いて音素アラインメントを行うモジュールを構築した．比較実験として，大人用音響モデルのみを用いた場合の評価を行った．実験結果を**表 2**，および**表 3** に示す．

表に示すように，大人用音響モデルのみを持つ場合，子供音声に対して正解アラインメント率の低下が見られる．一方，大人用音響モデルと子供用音響モデルの両方を用いることにより，正解アラインメント率の改善が得られた．大人音声，子供音声に対して 83 dBA の高騒音環境下でも 90%以上の性能が得られた．また 5 章に述べる本モジュールを実際に使用した被験者に対するアンケート調査の結果からも十分な性能が得られたことを確認した．

© 2008 Information Processing Society of Japan

表 2　大人用音響モデルのみを持つモジュールによる評価結果
Table 2　Performance of the system with adult acoustic models.

	騒音レベル（dBA）				
	63	68	73	78	83
大人音声（%）	99.16	99.16	98.99	98.66	96.81
子供音声（%）	97.82	97.32	97.15	93.96	86.41

表 3　大人用子供用の両方の音響モデルを持つモジュールによる評価結果
Table 3　Performance of the system with both adult and child acoustic models.

	騒音レベル（dBA）				
	63	68	73	78	83
大人音声（%）	99.16	99.16	98.99	98.83	96.98
子供音声（%）	98.83	98.66	97.15	95.97	90.60

4. アニメ調キーフレーム作成モジュール

デジタルアニメにおいて，リップシンクの実現は作品のクオリティを向上させる重要な要素の 1 つである．近年さまざまなリップシンクを含むフェイシャルアニメーションの検討が行われてきた．キーフレーム手法[5] や，人間の表情筋ベースによる物理シミュレーション法[6]，ビデオイメージを直接モーフィングに使用し，フェイシャルアニメーションを構築した研究[7]，モーションキャプチャを用いたアニメーション技術[8]，など多数の手法が提案されている．しかし，これらの手法の最終目標は，実際の人間の発話に近い，写実的なリップシンクアニメーションの実現を目的とした研究である．さまざまなスタイルのカートゥーンキャラクタに合わせてリップシンクアニメーションを作成する技術の検討は，ほとんど行われていない．

iFACe では任意の CG カートゥーンキャラクタに対して，音声分析結果から受け取った音素，音素継続長を調整し，キャラクタデザインに合うリップシンクを提供するため，Kawamoto らの手法[9] を用いた．この手法では 3 種類の様式化パラメータによりキーフレームを修正することで，対象の CG カートゥーンキャラクタに合うリップシンクに変換する．またキャラクタモデリング作業を軽減させるため，無表情（口を閉じたもの）とあ・い・う・え・おの 5 母音の口形状のジオメトリを用意し，これらのジオメトリ形状の重み付き線形和モデル（ブレンドシェープ法）を用いて，5 母音からさまざまな視覚素に対応した口形状の作成を行う．

本章では，5 母音から視覚素口形状の作成方法，キャラクタに合うリップシンク作成のための各種パラメータについて述べる．

表 4　母音–視覚素対応表
Table 4　Phonemes-vowels table.

	口形状（%）				
音素	A	I	U	E	O
a	100	0	0	0	0
i	0	100	0	0	0
u	0	0	100	0	0
e	0	0	0	100	0
o	0	0	0	0	100
g, k, ng	13	0	0	31	0
ch, s, sh, ts, z, zh, j	0	48	2	0	0
t	0	21	0	67	0
d, n	15	21	4	38	0
h, q	キーフレームなし				
f	0	31	0	0	0
b, m, p	0	0	0	0	0
r	0	0	11	68	0
w	0	0	41	0	0
silB	0	0	0	0	0
silE	0	0	0	0	0
-	直前の口形状の 80%				

4.1 母音から視覚素口形状へのマッピング

一般的に日本語音素に対する視覚素は十数種類程度である[18]．iFACe では Hiki らの手法[18] と Galatea FSM での音素–視覚素対応から経験的に**表 4** に示すような母音–視覚素対応表を作成した．視覚素の重みの計算は，Galatea FSM の標準顔モデルに含まれる，各視覚素に対応した口形状ジオメトリに対して，あ・い・う・え・お 5 母音のジオメトリの重み付き線形和から，最小二乗法を用いて算出を行った．

4.2 様式化パラメータ

音声分析結果をそのまま，5 母音口形状のキーフレームとして忠実に配置をしたときのカートゥーン発話アニメーションは，一般的に “うるさい” リップシンクアニメーションが作成される場合が多い．その理由の 1 つとして，カートゥーンのキャラクタスタイルが抽象的（アブストラクト）なキャラクタに対し，口形状が忠実に再現されることで違和感が生じるためである．そこで iFACe では次に示す，様式化パラメータ[9] を調整し，システムに使用するキャラクタに合うリップシンクアニメーションを構築している．

Cut パラメータ（0–100%）　“うるさい” 動きの口形状のキーフレームを間引くパラメータ

© 2008 Information Processing Society of Japan

で，各キーフレームの顔ジオメトリ頂点群の平均移動速度 V を算出し，最も V が大きいキーフレームから間引く．0%はすべてのキーフレームを残し，100%はすべてのキーフレームを消す．

Power パラメータ（0–100%）　音声パワーに応じて音声パワーが小さい場合は，キーフレームの重みを小さくする．0%は音声パワーの対応関係はなくし（各キーフレームの重みは一定量），100%は口形状のキーフレームの重みと音声パワーとが単純な比例関係となる．

Decay パラメータ（0–100%）　キーフレームの数は保ちつつ，キーフレームの重みを変化させて発話アニメーションの “なまけ” を表現する，0%は本パラメータの影響はなく，100%はなまけが最大になるよう設定されている．

4.3　様式化パラメータの適用例

様式化パラメータを適用した例を**図 4** に示す．図上段はアラインメント結果にそのまま口形状をあてはめた場合の 5 母音の混合重み（Baseline），中段は Baseline に対するキーフレーム位置での頂点移動速度の期待値，下段は Cut パラメータを用いて頂点移動速度を基にキーフレームを 1 つ取り除いた例が示されている．

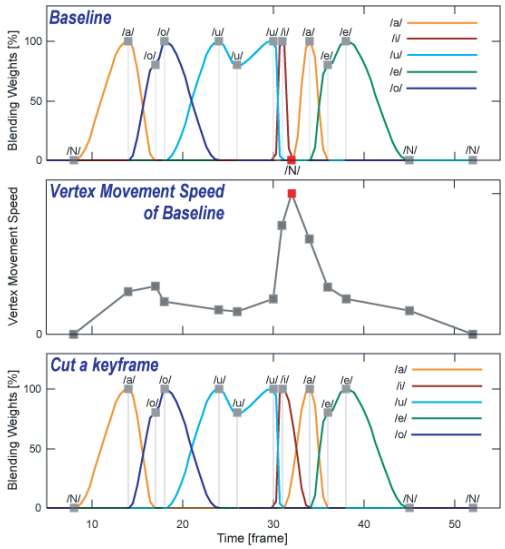

図 4　Cut パラメータによりキーフレームを 1 つ削除した例（上段：アラインメント結果にそのまま口形状をあてはめた場合の 5 母音の混合重み（Baseline），中段：Baseline に対するキーフレーム位置での頂点移動速度の期待値，下段：頂点移動速度を基にキーフレームを 1 つ取り除いた例）

Fig. 4　Example of removing a key-frame using a cut parameter (top: Key-frame sequence of baseline, middle: vertex movement speed of baseline, bottom: key-frame sequence applied a cut parameter for removing a key-frame).

灰色の矩形は，キーフレームの位置を示すものであり，/a,i,u,e,o/はそれぞれ日本語の 5 母音に対応する口形状「あ」「い」「う」「え」「お」を意味する．また，/N/は，口を閉じた無表情の顔モデルを示す．Cut パラメータでは，キーフレーム位置での頂点移動速度の期待値を基に，キーフレームの削除を行う．図上段のキーフレームに対応する頂点移動速度が図中段のものであり，この図において 33 frame 付近のキーフレーム/N/が最大の頂点移動速度を持つため，対応するキーフレームを図下段のように削除を行う．キーフレームの削除にともない，時間的に近接する頂点移動速度の期待値が変化するため，1 つキーフレームを取り除くごとに，頂点移動速度の期待値を再計算し，キーフレーム数が指定する割合になるまで，1 つずつキーフレームを取り除く．

また，**図 5** に，「こんにちは」と発話したときのキーフレームアニメーション生成結果を示す．

4.4　リアルタイムアニメーション生成モジュール

アニメ調キーフレーム作成モジュールから作成された，リップシンクデータを使ってリップシンクアニメーションを行う．iFACe はデジタルキャラクタ向けの声優体験システムであるため，セルアニメ調のトゥーンレンダリングが施された 3D キャラクタのアニメーションをリアルタイムに行う必要がある．

3D キャラクタのフェイシャルアニメーションを制御する場合，一般的に顔の中に仮想のボーンを埋め込み，ボーンを制御することでアニメーションを作成する方法（クラスタリング法）と，無表情キャラクタを変形させ，さまざまな表情（ターゲットシェープ）を作成し，ターゲットシェープの重み付き線形和モデルでアニメーションを作成する方法（ブレンドシェープ法）がある．クラスタリングの方が演算コストは少なく，一般的に速いため，ゲームなどリアルタイムの制御を行う際は，クラスタリング法を用いる場合が多いが，ボーンと皮膚との影響度のデザインを決定しなければならず，作業量が多い．

iFACe はターゲットシェープ法を採用し，無表情と 5 母音の口形状を最低用意すれば，トポロジの異なるキャラクタモデルでアニメーションできるよう考慮されている．演算コストの問題は，重み付き線形和の演算を GPU の頂点シェーダを用いることで，解決した．

リアルタイムアニメーション生成モジュールは C 言語＋OpenGL ARB＋NVidia Cg 言語で構築し，CPU：Intel Xeon 2.8 GHz，Graphics：NVidia GeForce 7600 のテスト PC で男女キャラクタ（男性：顔 43,916 ポリゴン，体全体：164,664 ポリゴン，女性（**図 6**）顔：3,736 ポリゴン，体全体 18,666 ポリゴン）でリフレッシュレート（60 Hz）と同等のフレー

© 2008 Information Processing Society of Japan

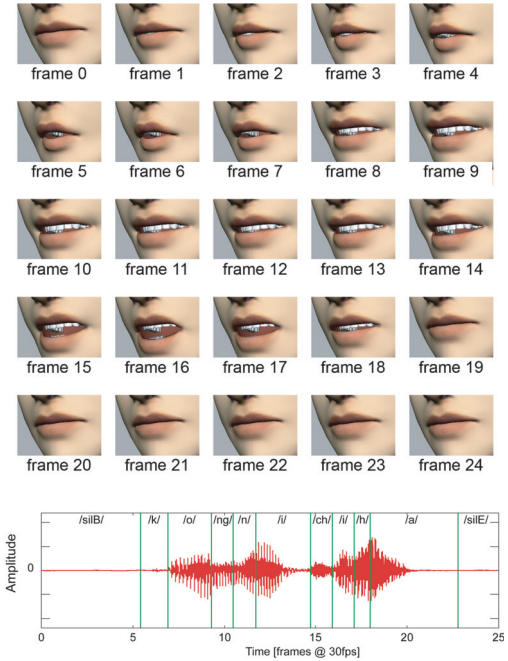

図 5　キーフレームアニメーション生成例（Cut = 10%，Power = 30%，Decay = 30%）
Fig. 5　Example of key-frame animation (Cut = 10%, Power = 30%, Decay = 30%).

図 6　描画テスト用女性キャラクタ
Fig. 6　3D lady character for display performance test.

ムレートで描画することができた．

5.　システム評価

iFACe の評価として，実際に iFACe を使ってもらい声優体験システムとしてのおもしろさ，リップシンクアニメーションの性能に関して主観評価実験を実施した．

5.1　評価用システムの実装

評価に使用した iFACe の立面図および実行風景を**図 7** に示す．PC は 3 台利用し，音声分析モジュールと，アニメ調キーフレーム作成モジュールは PC01，キャラクタ生成モジュールと FTP サーバは PC02 で動作させた．PC01 の構成は OS：Fedora Core4，CPU：Intel Pentium D 840（3.2 GHz），Memory：2 GB，PC02 の構成は OS：WindowsXP，CPU：Intel Xeon 2.8 GHz，Graphics：NVidia GeForce 7600，Memory：2 GB である．参加者はまず，PC00：タッチパネル NotePC00 を用いて，台詞を選択し，音声を収録する．収録した音声は FTP サーバに格納され，分析タスクの開始コマンドをシステム操作モジュール

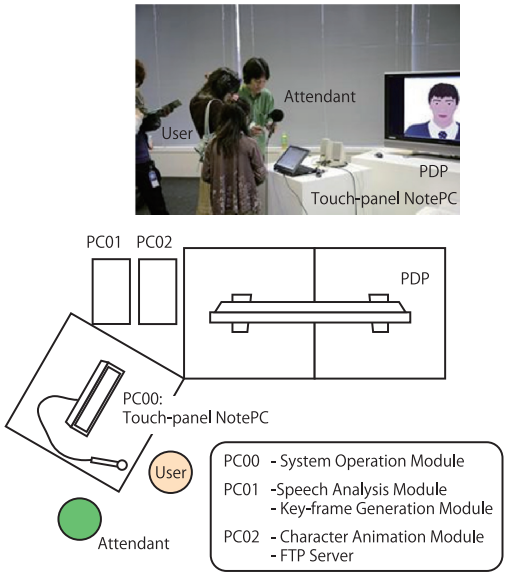

図 7　システムの実装（写真は実行風景）
Fig. 7　System implementation.

© 2008 Information Processing Society of Japan

から音声分析モジュールへ渡す．音声分析モジュールで音声データから音素・音素継続長を分析し，アニメ調キーフレーム作成モジュールで，リップシンク用キーフレームの作成を行った．この際に利用した，様式化パラメータはリップシンク用キャラクタのスタイルに合うよう経験的に，Cut = 10%，Power = 30%，Decay = 30%とした．最後にキャラクタ生成モジュールで，キャラクタと音声の同期出力を行った．映像出力は大勢の参加者が視聴できるよう，42 インチ PDP を用いた．使用したキャラクタは男女 2 体（男性：顔 9,020 ポリゴン，体全体：45,316 ポリゴン，女性：顔 8,868 ポリゴン，体全体 67,330 ポリゴン），システム操作モジュールで参加者が自由に選択することができるように，操作ウインドウにボタンを配置した（図 2）．

5.2　実験評価環境

日本科学未来館にて，2006 年 5 月 3 日から 7 日まで，（独）科学技術振興機構主催の「予感研究所」で iFACe を展示した．展示会場は iFACe だけではなく，音を発生する展示物も多数あり，展示時の雑音は常時 70〜80 [dBA] 程度であった．主な体験者は親子連れの参加者で，5 日間延べ 1,500 回以上の発話アニメーションを生成した．展示最終日を中心に，iFACe の性能評価アンケートを配布し．有効回答数が 84 であった．

5.3　エンタテインメントシステムとしての有効性に関する評価

iFACe を体験してもらった参加者に次の質問を行った．

質問 1「おもしろい」

質問 2「アミューズメントシステム，ゲームソフトにあったらまた遊んでみたい」

被験者は 5 段階（[5]：非常にそう思う，[4]：ややそう思う，[3]：どちらでもない，[2]：あまりそう思わない，[1]：全然そう思わない）で回答した．また質問 1，2 の選択肢のほか，質問 1' 具体的にどのようなところがおもしろい・おもしろくないのか，質問 2' どのようなキャラクタで遊んでみたいか自由記述による質問を行った．

評価結果を**図 8** に示す．質問 1 から 99%の回答者（平均スコア：4.64）が，本システムをおもしろいと判断した（評価 4 以上）．また質問 2 から 74%の回答者（平均スコア：4.04）が，ゲームとして声優体験システムをまた遊んでみたいと示した（評価 4 以上）．質問 1' では，「声優になった気分」「自分の声でキャラクタがしゃべるところ」に好評価をいただき，また「（台詞が）つっかかってもちゃんと表現される」というコメントがあり，言いよどみに関しても，アニメーションクオリティを保つことができたことが分かった．質問 2' では有名ゲーム・漫画キャラクタや，芸能人のキャラクタで喋ってみたいとのコメントがあった．以上の回答結果から iFACe がエンタテインメントシステムとして非常に有効なコンテンツであることが分かる．

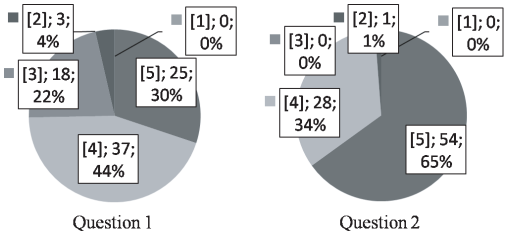

図 8　システムの有効性の評価（項目; 人数; パーセンテージ）
Fig. 8　Effectiveness evaluations for entertainment system.

5.4　リップシンクアニメーションの性能評価

リップシンクアニメーションに関する評価を行うため，以下の質問を行った．

質問 3「キャラクタが喋っていたときの口は自然に動いていましたか？」

被験者は 4 段階（[4]：はい，[3]–[1] いいえ（[3]：わずかに不自然であるが気にならない程度，[2]：すこし不自然，[1]：不自然））で回答した．

質問 4「キャラクタの口の動きと声が違和感なくちゃんと合っていましたか？」

被験者は 4 段階（[4]：はい，[3]–[1] いいえ（[3]：わずかに違和感があるが気にならない程度，[2]：すこし違和感がある，[1]：違和感がある））で回答した．

質問 5「アニメーションはなめらかに動いていましたか？」

被験者は 4 段階（[4]：はい，[3]–[1] いいえ（[3]：わずかにぎこちないが気にならない程度，[2]：すこしぎこちない，[1]：ぎこちない））で回答した．

質問 6「アニメーションができるまでの待ち時間は十分に短かった」

被験者は 5 段階（[5]：非常にそう思う，[4]：ややそう思う，[3]：どちらでもない，[2]：あまりそう思わない，[1]：全然そう思わない）で回答した．

評価結果を**図 9** に示す．質問 3 では 62%の回答者が自然なリップシンクアニメーションであると示され，[3] 気にならない程度までを加えると 86%であった（平均スコア：3.46）．

質問 4 では 71%の回答者がキャラクタの口の動きと音声とがリップシンクの精度に関して問題ないと示し，[3] 気にならない程度までを加えると，90%であった（平均スコア：3.58）．この結果から，音声分析モジュールが精度良く，音素アラインメントが行われたことが分かる．

質問 5 では 61%の回答者がなめらかに発話アニメーション動作したと評価し，[3] 気にならない程度を加えると 88%であった（平均スコア：3.48）．質問 3，5 からアニメ調キーフ

© 2008 Information Processing Society of Japan

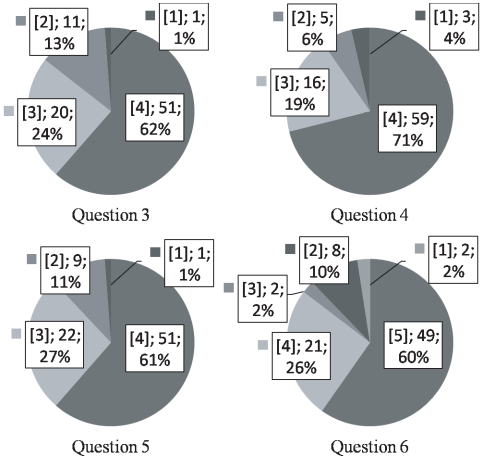

図 9　リップシンクアニメーションの性能評価（項目; 人数; パーセンテージ）
Fig. 9　Performance evaluations for lip-sync animation.

レーム作成モジュールによる様式化パラメータを用いたため，良好な結果を得ることができたと考えられる．

質問 6 では 86%の回答者（平均スコア：4.30）が音声入力からアニメーション出力までの時間は短かったと示した．しかし [2] 待ち時間が短いとは思わない参加者が 10%という結果となった．待ち時間の大部分は音声分析処理に使用しており，また発話処理のたびに音声分析モジュールはデータサイズの大きい，各音響モデルの読み出しを行っているために遅延が発生していることが考えられる

以上の結果から，iFACe のリップシンククオリティは参加者に満足のゆくアニメーションを提供していることが分かる．また今回のような，雑音環境下での運用も問題ないことが示された．

6. 議　　論

評価実験時，参加者から決められた台詞ではなく自由に発話した音声でリップシンクアニメーションの作成ができないか，に関するコメントをいただいた．その参加者に対してプロトタイプとして用意した自由発話に対応した iFACe を体験してもらった．本章では．自由発話モードによる音声分析モジュールの性能評価実験および考察を述べる．また，アンケート調査で本システムに関するコメント（自由記述）から，iFACe のシステム拡張性に関して述べる．

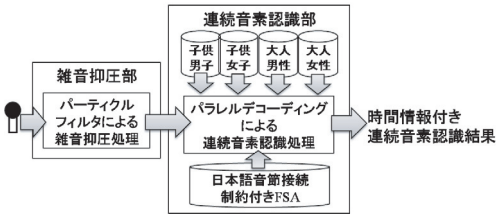

図 10　自由な発話に対応した音声分析モジュールの構造
Fig. 10　Overview of speech analysis module for unlabelled utterance.

6.1　自由な発話に対応した音声分析モジュール

アンケート結果から，「携帯電話の伝言サービスに取り入れて，（サービスを利用した）人のアニメーションで声を出す」サービスがあるとおもしろいとの意見があり，今後，自由発話モードを用いることで iFACe の利用範囲が広がると考えられる．本節では，発話内容をあらかじめ決めることなく，ユーザの発話した自由な発声（発話内容未知）に対してリップシンク動画像を自動生成するための音声分析モジュールの評価を行った．本音声分析モジュールの構造を**図 10** に示す．

3 章で述べた既知の発話内容のための音声分析モジュールと同様に，2 つの処理部から構成される．マイクより入力された音声波形は，雑音抑圧部により背景雑音が抑圧される．連続音素認識部では，日本語音節接続の制約を表現した FSA 型言語モデルを用い，入力された音声に対して Viterbi アルゴリズムにより最尤の音素列，および個々の音素の開始時刻が推定される．この FSA には，日本語音節の基本構造である母音＋子音の制約が記述されている．HMM によりモデル化された子供用音響モデルと大人用音響モデルの両方を用い，パラレルデコーディングにより認識が行われる．これらの音響モデルは，3 章で述べた音声分析モジュールで用いたモデルと同じである．

6.1.1　実 験 条 件

音声分析モジュールの性能を評価するため，3 章で用いた評価データに対する音素アラインメント実験を行った．大人 6 名（男性 3 名，女性 3 名，各話者 10 文，計 60 文）と，子供 6 名（男子 3 名，女子 3 名，各話者 5 文，計 30 文）である．評価尺度として，検知限による正解音素アラインメント精度を用いた．評価対象とした音素は母音のみである．本実

© 2008 Information Processing Society of Japan

表 5　ベースラインモジュールの性能評価結果
Table 5　Performance of the baseline module.

	騒音レベル（dBA）				
	63	68	73	78	83
大人音声（%）	82.67	75.99	63.47	38.41	17.75
子供音声（%）	59.71	52.82	45.09	29.65	11.48

表 6　提案モジュールの性能評価結果
Table 6　Performance of the proposed module.

	騒音レベル（dBA）				
	63	68	73	78	83
大人音声（%）	86.43	84.55	78.91	68.06	46.76
子供音声（%）	65.76	63.88	58.87	46.14	35.28

験により得られる開始時刻付き音素列は，正解音素列と異なる可能性があり，挿入，削除，置換誤りなどを含む．正解音素列との比較には，動的計画法により最大の正解音素アラインメント精度が得られるように，正解音素と認識音素の間のマッピングを決定した．正解音素アラインメント精度（%）は，$(N-I-D-S)/N$（N は音素数，I は挿入誤り数，D は削除誤り数，S は置換誤り数）により計算される．

6.1.2　評価実験結果

表 6 に，本音声分析モジュールの正解音素アラインメント率を示す．比較として，**表 5** に雑音抑圧処理なし，大人用音響モデルのみのベースラインモジュールの性能を示す．表に示すように，本システムは，ベースラインモジュールと比較して，大人，子供音声，およびすべての騒音レベルにおいて高い正解音素アラインメント精度が得られた．

しかしながら，本提案システムの評価結果である表 6 と，発話内容を既知として評価を行った表 3 を比較して，大人音声の場合 63 dBA で約 13%の低下，83 dBA で約 50%の低下が見られる．今後は，発話タスクを限定した単語 n-gram 型言語モデルなどの利用により，比較的自由な発話に対する音素アラインメント性能の改善を行う予定である．3 章で述べたシステムは，あらかじめ与えられた発話内容とまったく同じ発声に対して音素アラインメントを行うシステムである．それに対して，単語 n-gram 型言語モデルを用いることにより，タスク内における自由発話に対して，本システムで用いた日本語音節接続制約型言語モデルよりも高い音素アラインメント精度が得られると考えられる．

6.2　システムの拡張性

参加者からのコメントを含む自由回答結果から，iFACe のシステム拡張性について述べる．

6.2.1　CG キャラクタに関する拡張性

評価実験では男女 2 体のキャラクタを参加者の希望に応じ，表示させていた．アンケート調査結果では「自分の好きなキャラクタ」や「有名マンガ・アニメのキャラクタ」，「ハリウッド俳優」，「芸能人」で遊んでみたいとの意見があった．現在の iFACe は 3DCG ソフトウェアで作成した無表情の顔形状と，母音口形状のポリゴンデータ，およびテクスチャデータがあれば，iFACe の登場キャラクタとして追加可能である．また，キャラクタデザインに応じて，4 章で述べた様式化パラメータを調整することで，キャラクタに合うアニメ調キーフレームを作成することができる．調整には iFACe の描画モジュールを用いることで，リアルタイムに発話アニメーションが確認できるため，短時間で行うことができる．なお，現状のシステム（2 キャラクタ，8 台詞）の調整では，30 分程度で完了することができた．現在はキャラクタに応じて，様式化パラメータを調整していたが，将来的に発話単位でパラメータが自動調節できる機能を付加させることで，より自然な発話アニメーションが生成できると考えられる．またアンケートには「もっとキャラクタが豊富だといい」との意見があった．評価では 2 体のキャラクタのみ表示していたが，GPU メモリの上限まで複数のキャラクタを登場させることが可能である．

そのほか「顔の表情も同時に動くともっと良い」，「口以外の部分も同時に動くようになるともっと滑らかになると感じると思う」との意見があった．評価実験時は表情を考慮せず，表情が必要な場合は，別途操作画面に用意した表情ボタン（喜び，怒り，悲しみ，驚き）を押すことで，表情が出力される機能のみを用意した．今後の機能拡張として，音声分析から発話タイミング情報だけではなく，感情推定手法を用いて，自動的に表情を出力し CG キャラクタの表現力を高めることができると考えられる．

6.2.2　他分野への利用

アンケート結果から「教育・医療現場などで活かすことができれば良いと思う」や「英会話の発話レッスンにも使いたい」など，エデュテイメント性のある語学教育[19] へのアプリケーション化について意見があった．iFACe は CG キャラクタ作成作業を軽減させるため，アニメーションに必要な口形状は母音の「あ・い・う・え・お」のみに限定していた．教育・医療分野へ応用化する際は，正確な口形状を必要とするため，視覚素に対応した口形状すべてを作成する必要があるが，音素・視覚素対応表を更新することで，利用することが可能である．また日本語以外の他言語リップシンクアニメーションは，音声分析モジュールの変

© 2008 Information Processing Society of Japan

更，音素・視覚素対応表の変更を行うことで，現在の iFACe のワークフローを変えず，利用ができると考えられる．

また「携帯端末を使ったアバタによる留守番メッセージ自動読み上げサービスの利用」のモバイルコンテンツへの応用化に関しても意見があった．PC に搭載できるグラフィックスアクセラレータを利用することはできないが，近年 GPU を搭載した携帯電話が販売され，数百〜数千ポリゴン＋低解像度テクスチャを使ったモバイル用キャラクタ生成モジュールを構築すれば iFACe のワークフローを利用することで実現できると考えられる．

7. ま と め

専門スキルを必要とせず，参加者の音声とデジタルアニメキャラクタとのリップシンクを体験することができるエンタテインメントシステム向け，声優体験システム iFACe を提案した．日本アニメの一般的な制作手法であるアフレコ作業ではなく，参加者の音声を分析し，自動でキャラクタの発話アニメーションの作成を行うプレスコを用いることで，子供でも簡単に体験できるように考慮した．アミューズメント施設での運用を想定し，雑音環境下でも安定に音素・音素継続長を推定する音声分析モジュールを用いることで，客観評価から大人音声，子供音声に対して 83 dBA の高騒音環境下でも正解アラインメント率が 90%以上の性能が得られた．また主観評価実験から 90%の回答者からリップシンクの同期のズレが気にならない結果を得ることができた．さらに，アニメ調キーフレーム作成モジュール・リアルタイムアニメーション生成モジュールを用いることで，なめらかな発話アニメーションを生成することが可能となった．システム評価から，iFACe がエンタテインメントシステムとして有効であることを示せた．

謝辞　本研究の一部は科学技術振興機構（JST）の戦略的基本研究推進事業（CREST）支援によるものである．

参 考 文 献

1) Thomas, F. and Johnston, O.: The Illusion of Life: Disney Animation, *Disney Editions* (2005).
2) 中野渡昌平，宇賀神岳史，田中義也：特開 2008-022979 (2008).
3) 中西宗博，池田宜史：特開 2006-346284 (2006).
4) 河原達也，李　晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol.20, No.1, pp.41–49 (2005).
5) Pighin, F., Hecher, J., Lischinski, D., Szeliski, R. and Salesin, D.H.: Synthesizing realistic facial expressions from photographs, *Proc. 25th Annual Conference on Computer Graphics and Interactive Techniques* (*SIGGRAPH'98*), pp.75–84, (1998).
6) Waters, K.: A muscle model for animation three-dimensional facial expression, *Proc. 14th Annual Conference on Computer Graphics and Interactive Techniques* (*SIGGRAPH'87*), pp.17–24 (1987).
7) Ezzat, T., Geiger, G. and Poggio, T.: Trainable videorealistic speech animation, *Proc. 29th Annual Conference on Computer Graphics and Interactive Techniques* (*SIGGRAPH 2002*), pp.388–398 (2002).
8) Scott, R.: Sparking life: Notes on the performance capture sessions for the lord of the rings: The two towers, *SIGGRAPH Comput. Graph.*, Vol.37, No.4, pp.17–21 (2003).
9) Kawamoto, S., Yotsukura, T., Anjyo, K. and Nakamura, S.: Efficient lip-synch tool for 3D cartoon animation, *Computer Animation and Virtual Worlds*, Vol.19, Issue 3–4, pp.247–257 (2008).
10) Yotsukura, T., Morishima, S. and Nakamura, S.: Model-based talking face synthesis for anthropomorphic spoken dialog agent system, *ACM Multimedia 2003*, pp.351–354 (2003).
11) 四倉達夫，藤井英史，森島繁生：サイバースペース上の仮想人物による実時間対話システムの構築，情報処理学会論文誌，Vol.40, No.2, pp.677–686 (1999).
12) Fujimoto, M. and Nakamura, S.: A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging, *IEICE Trans. Information and Systems*, Vol.E89-D, No.3, pp.922–930 (2006).
13) 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 2005 年秋期研究発表会講演論文集，2–P–22 (2005).
14) 實廣貴敏，松田繁樹，藤本雅清，Wolfgang Herbordt，堀内俊治，中村　哲：ATR における日本語音声認識の評価—日本語音響モデル，日本音響学会 2006 年度春季研究発表会，1–P–21, pp.185–186 (2006).
15) Center for Integrated Acoustic Information Research (online). available from http://db.ciair.coe.nagoya-u.ac.jp/eng/dbciair/dbciair2/kodomo.htm (accessed 2008-09-17)
16) 赤井田卓郎，岡田清孝，黒住幸一，林　俊一，深谷崇史：リップシンク—映像と音声のタイミング，NHK 技研だより，1997 年 5 月号 (1997).
17) 河井　恒，戸田智基：波形接続型音声合成のための自動音素セグメンテーションの評価，信学技報，SP2002–170, pp.5–10 (2003).
18) Hiki, S. and Fukuda, Y.: Characteristics of the Mouth Shape in the Utterance of Japanese, *ASJ Trans. Comm. on Speech Res.*, S76–49, pp.1–8 (1977).

© 2008 Information Processing Society of Japan

19) Massaro, D.W.: A computer-animated tutor for language learning: Research and applications, *Advances in the spoken language development of deaf and hard-of-hearing children*, Spencer, P.E. and Marshark, M. (Eds.), Oxford University Press, pp.212–243 (2006).

(平成 20 年 3 月 26 日受付)
(平成 20 年 9 月 10 日採録)

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．

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．

松田　繁樹（正会員）

1997 年帝京大学情報科学科卒業．1999 年北陸先端科学技術大学院大学博士前期課程修了．2003 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，同研究所上級研究員，および（独）情報通信研究機構専門研究員．雑音環境下における音声認識に関する研究に従事．電子情報通信学会，日本音響学会各会員．

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．

© 2008 Information Processing Society of Japan