社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＨＩ－109　(9)
2004－SLP－ 52　(9)

2004／7／16

アート・エンターテインメントにおける音インタフェース

江渡　浩一郎†

音声認識，音声合成などの音声処理技術はすでに成熟した技術となったが，まだ一般に普及したとは言えない状況である．コンピュータとのインタフェースに音声を用いる音声インタフェースの応用事例として，本稿では，非日常的な状況，アートやエンターテインメントといった分野における音声処理技術の応用事例についての動向をまとめるものである．音声認識，音響処理，歌唱合成などを応用した事例において，近年注目すべき作品が出てきている．

Voice Interface in the field of Art and Entertainment

Kouichirou Eto†

Although sound processing technologies, such as speech recognition and speech synthesis, are matured, they are not in widespread use yet. I explore the current movement of the applications in the field of Art and Entertainment, as application of Voice Interface, using voice for man-machine interface. Recently, there are some remarkable works in the applications using speech recognition, acoustic analysis and singing voice synthesis.

1. 音声インタフェースの現状

音声認識，音声合成などの音声処理技術は古くから研究・開発が進められており，現在ではすでに成熟した技術になってきたと言えるだろう．しかし，コンピュータに対するインタフェースとして音声処理技術を用いることは，まだ一般的になったとは言えない．音声処理技術を用いたインタフェース（音声インタフェース）が一般に普及していない理由は認識精度の問題などがあげられるが，一つにはまだ適切な応用事例がみつかっていないという理由も考えられるのではないだろうか．一般に技術開発は，広く日常的に利用されるような実用化を目標とするが，本稿ではその日常的な文脈における実用化という目標を一旦離れ，アートやエンターテインメントといった非日常的な分野における応用事例に目を向けてみる．その応用事例から，現在はまだ探索されていない新しい利用分野に目を向けることができるかもしれない．

音声インタフェースを大きく二つに分類すると，音声入力と音声出力になる．音声入力は，人間が発話する自然言語を理解する音声認識技術が一般的だが，音声の音量や音高，スペクトル分析などを行う音響処理技術もその一手法である．音声出力については，サンプリング化された音声を再生する技術，MIDI 音源のように楽音を合成する技術がすでに広く普及しているが，人の声を合成する音声合成技術はまだあまり一般に普及しているとは言えない．音声合成技術の発展形として歌声の合成，歌唱合成技術について近年いくつかの応用事例が出てきている．本稿では，アート・エンターテインメント分野における，それぞれの技術についての応用事例を紹介する．

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

2. 音声入力を応用したアート作品

2.1 「Messa di Voce」

Tmema（Golan Levin と Zachary Lieberman の二人からなるグループ）による「Messa di Voce」[1)2)]（イタリア語で”placing the voice”という意味）という作品は，「もし声を見ることができるとしたら，どんな風に見えるだろう」というアイデアを元にした，二人のボイスパフォーマーによって繰り広げられるパフォーマンスである．2003 年 9 月 7 日に，オーストリアのリンツにおける Ars Electronica Festival において初演された．ステージ上には高さ 3m，横 8m のスクリーンが設置され，二人のボイスパフォーマーが 30〜40 分程度のパフォーマンスを繰り広げる．その様子を画像解析し，頭の位置を認識する．それぞれのパフォーマーの声を音響処理し，その結果を抽象的な

図 1　「Messa di Voce」Bounce のシーン．顔の位置が認識され，口の位置から玉が出てくるように表示される．

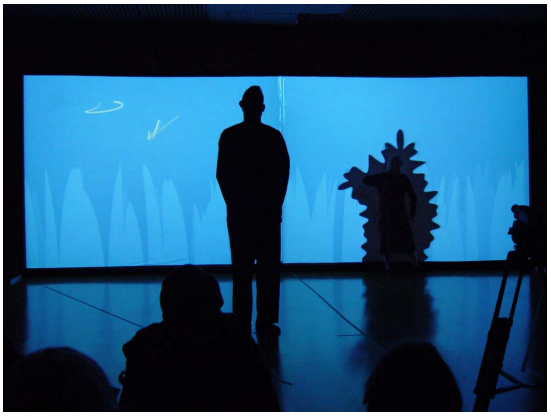

図 2　「Messa di Voce」Insect のシーン．右側の女性の輪郭が，黒いかたまりのように表示されている．

形態として表示する．パフォーマンスは全部で 12 個のシーンから成っており，いくつかを解説する．

Bounce　パフォーマーが一つ一つの音声を区切って発音すると，それが玉のような形として表示される．（図 1）画像認識によってパフォーマーの顔の位置が認識され，あたかも口の位置から玉が出現したかのように表示される．玉の動きは物理シミューレションによって計算されるが，画面上の重力は逆転しており，上の方に玉が溜っていく．ある一定時間経つと突然重力が反転し，玉は落ちてくる．個々の玉には発話の際の音声が録音されており，玉が落ちて地面に当る瞬間に再生される．玉が雨のように地面に落ちるとき，その音が一斉に再生される．

Insect　画像解析で人物の輪郭を抽出し，人物が黒い影の固まりになったように表示される．（図 2）その人物の声の音高によって，輪郭を波立たせる．

PitchPaint　画面上にピッチによって方向が変わる線が表示される．（図 3）音量が線の太さに，音高

図 5　「n-Cha(n)t」天井から吊り下げられた 7 台のモニター．エージェントは，お互いに音声認識・音声合成を行う．

が線の方向に対応している．うまく音高を制御すると，任意の形を描くことができる．線が画面上で交差すると，その囲まれた領域が任意の色で塗られる．「シュー」というノイズを出すと，線が消去される．

2.2　「RE:MARK」「Hidden Worlds of Noise and Voice」

同じ作者による一つ前の作品が「RE:MARK」と「Hidden Worlds of Noise and Voice」[3] である．「RE:MARK」では，観客がマイクに向かって声を入力すると，その声を音高，スペクトル重心，メルケプストラム係数などの音響パラメータによって分析し，結果を抽象的な形態として表示する．音素として認識できた場合は，アルファベットとして表示する．「Hidden Worlds of Noise and Voice」では，観客は HMD を装着し，マイクに向かって声を出すと，その声に応じた抽象的な形態が仮想空間中に生成され，動きまわる．（図 4）

2.3　「n-Cha(n)t」

David Rokeby による「n-Cha(n)t」[4]（エヌ・チャント）という作品は，7 台のエージェントが相互に音声認識・音声合成しあうことで，詩的な会話が生成されるという作品である．薄暗い展示空間の中で，天井から 7 台の液晶モニターが吊り下げられている．それぞれのモニターには「耳」が映っており，そのエージェントの状態を示している．（図 5）観客がマイクにむかって話しかけると，その言葉は外部からの信号としてそれぞれのエージェントへ送られる．入力された言葉を元に連想を行い，その言葉をテーマとした一連の詩を生み出し，音声合成によって朗読する．その詩はエージェント間にも送られ，相互に影響を与えあう．だれも観客がいない空間では，調和のとれた歌を歌うようになる．2001 年に Canada の Banff Centre

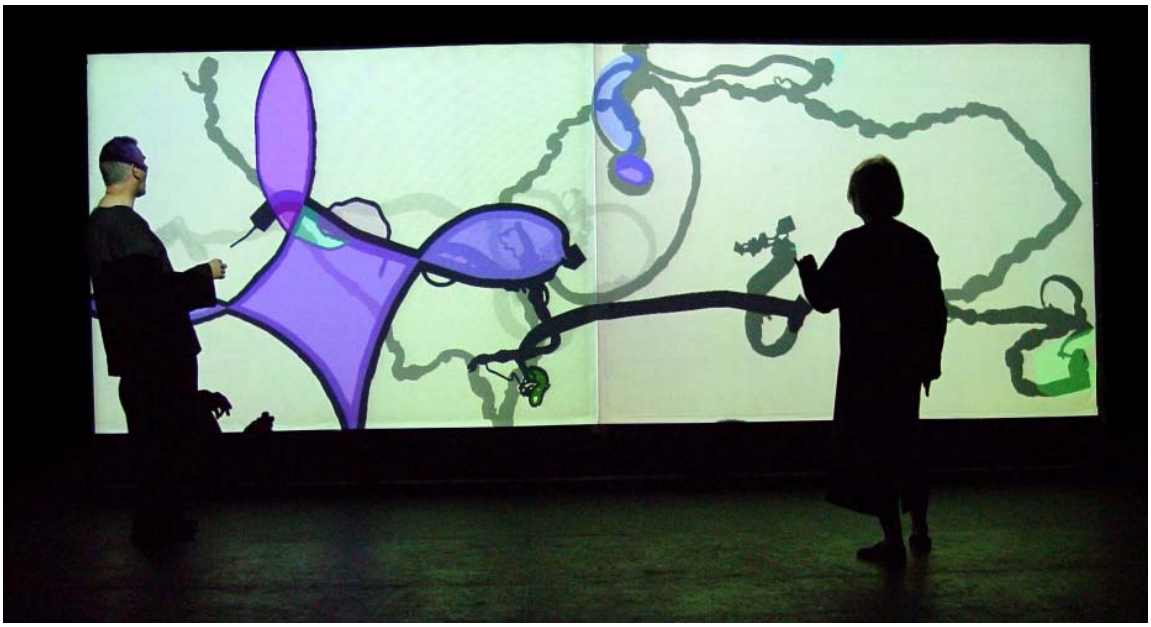

図 3　「Messa di Voce」PitchPaint のシーン

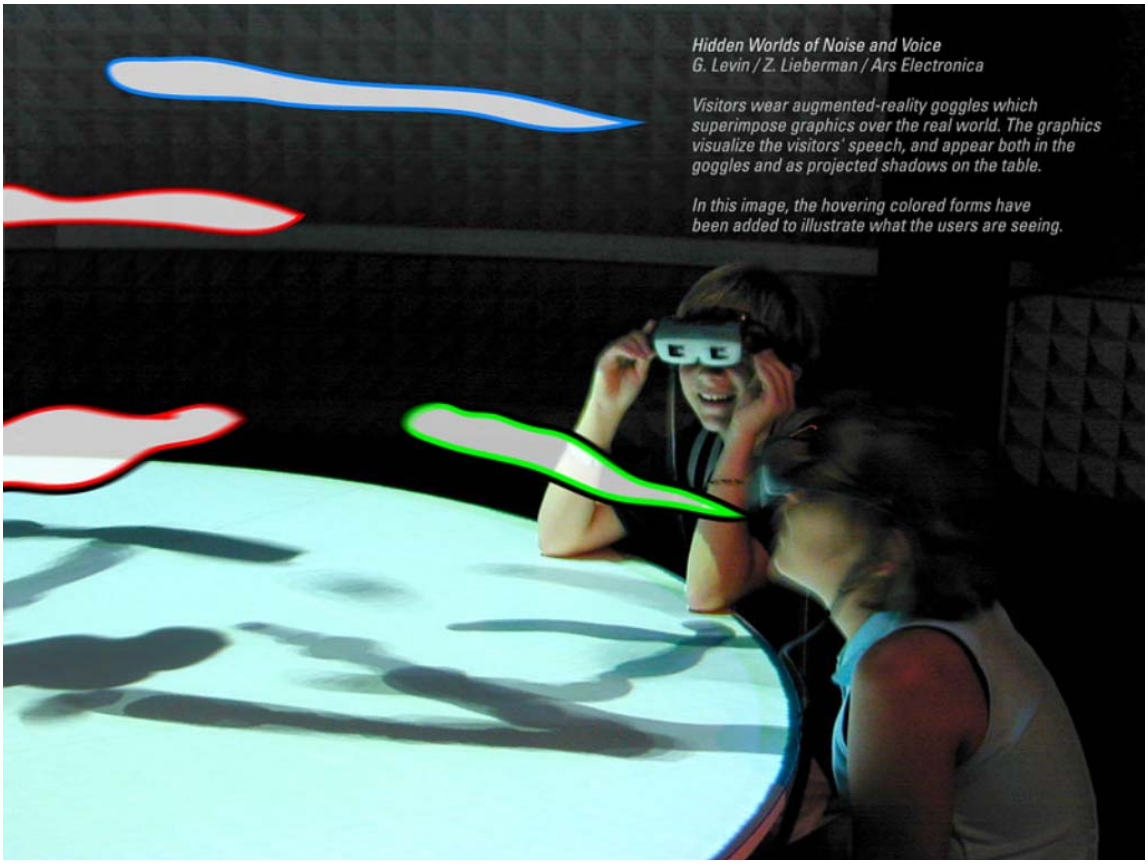

図 4　「Hidden Worlds of Noise and Voice」

図 6　「およぐことば」マイクにむかって話した声が，水面を漂う．

for the Arts における「Computer Voices / Speaking Machines」展の一作品として展示された．David Rokeby はこの作品で 2002 年 Ars Electronica 賞の Interactive Art 部門グランプリを受賞した．

2.4　「およぐことば」

師井聡子，笹田晋司，柴田良二による「およぐことば」[5] という作品は，音声認識を用いて言葉遊びをする作品である．観客がトランペット状のマイクにむかって話しかけると，音声認識され，文字となって水面を漂いはじめる．(図 6) 魔法の杖を使って，文字をかき混ぜることもできる．杓でその文字を持ちあげると，その文字を表す形態へと変化する．例えば「あ」という文字は「あり」の形に姿を変えるかもしれない．沖縄市にある「沖縄こども未来ゾーン」にて常設展示されている．

2.5　「フキダマリ」

「フキダマリ」[6] は，マイクから入力された音声を，フキダシのような形で表現する作品である．観客はマイクにむかって言葉を話しかけると，音声はマンガにおけるフキダシのような抽象的な形として壁面上に表示される．(図 7) その形態は声の音量，音高，抑揚などによって決定される．そのフキダシを叩くと，フキダシの中に録音された音声が再生される．フキダシは壁面を漂い，時間がたつにつれて徐々に音声は変化していく．北京にある「ソニーエクスプロラサイエンス」にて常設展示されている．

2.6　「FMS+ThirdEar」

exonemo による「FMS (FragMental Storm)」[7] は，任意のキーワードを元にインターネットで検索を行い，関係のある言葉や画像を画面上でランダムに再構成するという作品である．(図 8) そのインスタレーション版である「FMS+ThirdEar」は，カフェの店員にマイクを付け，そこでかわされている会話を音声認識し，その単語を元に検索を行う．

図 7　「フキダマリ」マイクにむかって話した声が，フキダシとして表示される．

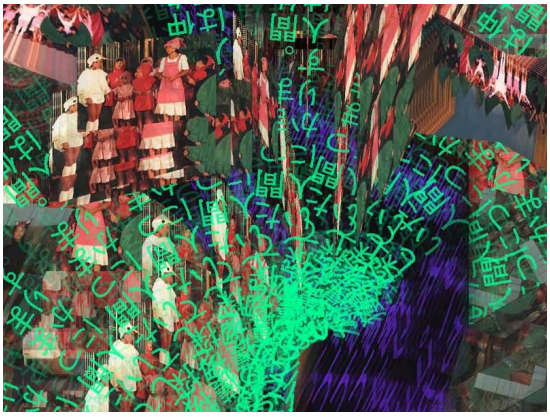

図 8　「FMS」インターネットから検索された言葉や画像がランダムに表示される．

2.7　「D SYSTEM」

フランスと日本の音楽家 10 人によるグループ PACJAP[8] は，コンピュータを介した新しい形のパフォーマンスを探究するための音楽装置「D SYSTEM」を制作した．この装置は，入力した音の様々なパラメータを，他のパラメータに変換して出力するというものである．例えばあるシーンでは，詩人が歌う歌声が入力となり，その音高が音量に，音量が音高へと入れ替えて音声出力される．コンピュータを介した音声の入出力のあり方に疑問を呈するという試みである．図 9 に概念図を示す．

2.8　「SendMail」

「サクソフォン，ピアノとコンピューターのための SendMail」[9][10] は三輪眞弘によるパフォーマンス作品である．サクソフォンによって演奏される音声が，MAX/MSP によって音高解析され，その結果はキーボードからの文字入力へと変換される．二音が，一打鍵にマッピングされる．サクソフォンだけによる文字

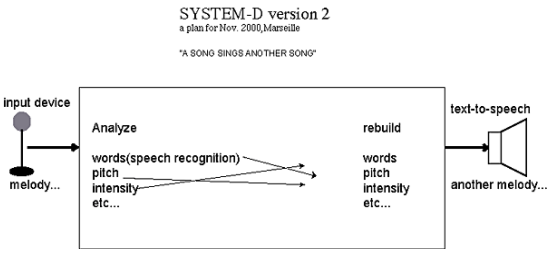

図 9 「D System」概念図

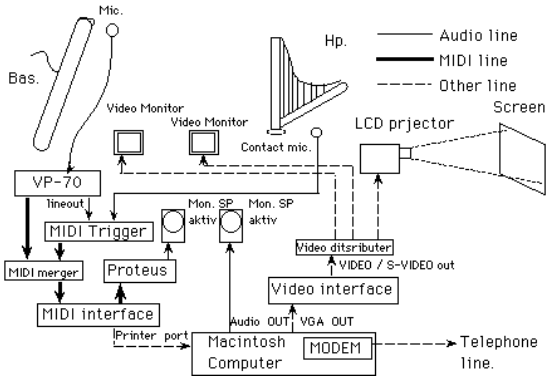

図 10 「SendMail」構成図

入力によって，モデムによるダイアルアップ接続から始まり，インターネットに接続し，コマンドラインからメール送信を行う．図 10 に構成図を示す．これは現在当たり前の物となっているキーボードからの入力というインタフェースに対して，音声による入力というオルタナティブを提示する作品である．

3. 音声入力によるエンターテインメント作品

3.1 音響認識技術を応用したゲーム

音声認識技術を応用したゲームとしては，1998 年に発売された「ピカチュウげんきでちゅう」，1999 年に発売された「シーマン」が先駆的な存在である．また，PC 用ゲームにおいて,「ぼくは航空管制官」「電車で Go!」のように音声認識によりコマンド入力を行う事例がある．

3.1.1 「オペレーターズ・サイド」

2003 年にソニーコンピュータエンターテインメント社 (SCEI) から発売された画面上のキャラクターに声で指示を出し謎を解決していくアドベンチャーゲーム．スキャンソフト社[11] 製の音声認識ミドルウェアを使用している．

3.1.2 「犬とあそぼう dogstaion」

2003 年にコナミより発売された仮想の犬を育てるゲーム．犬に話しかける言葉によって，成長の様子が変わってくる．音声認識ミドルウェアとして，ソニー「S-FORCE」[12] を使用している．

3.1.3 「デカボイス」

2003 年にアクワイアより発売された刑事となって事件を解決するアドベンチャーゲーム．音声入力を使って相棒を誘導し，聞き込み調査，取調べなどを行う．

3.2 音響処理技術を利用したゲーム

3.2.1 「ビブリボン」

1999 年に SCEI より発売されたリズムアクションゲーム．普通の音楽 CD から音楽トラックを読み込み，音響処理によってリズムを認識し，アクションの舞台を自動生成する．任意の音楽 CD をゲームの舞台として利用することができる．

3.2.2 「アフレコ！」

2002 年にナムコが開発したアニメの登場人物になりきって演技をするゲーム．画面ではアニメの一場面が上映され，その中の一人の登場人物になりきって，マイクにむかって声で演技する．その音声が録音され，音量，正確さ，抑揚などが音響処理によってパラメータ化され，スコアが与えられる．アーケイド用のゲームとして開発され，ロケテストまでは行われたが，市場には出ることはなかった．

3.2.3 「しばいみち」

2003 年に SCEI から発売された画面上の登場人物になりきって声で演技をするゲーム．画面上のセリフをマイクに向かってしゃべり，演技力を競う．演技力は「正確さ」「声量」「感情」の 3 つの要素で判定される．

3.2.4 「ドンキーコンガ」

2003 年に任天堂より発売されたリズムアクションゲーム．付属のタル型の入力装置を叩いてリズムを入力する．その装置にはマイクもついており，音声入力による手拍子も認識する．

4. 歌唱合成技術を利用した作品

4.1 「モジブリボン」

2003 年に SCEI より発売されたリズムアクションゲーム．音声合成を用いてラップを自動生成する．インターネットから受信したメールを，ラップとして歌ってくれる機能もついている．音声合成ミドルウェアとして NTT アイティ「Hipervoice」[13] を使用している．

4.2 「くまうた」

2003 年に SCEI より発売された，くまに演歌を教えるゲーム．くまは自動作詩・作曲によって，演歌を生成し，歌唱合成によって歌う．インターネットから受信したメールを演歌として歌ってくれる機能もついている．

音声合成ミドルウェアとしてアニモ「FineSpeech」[14]を使用している．

4.3 「ミクロ楽団」

PHS を装着した PDA をデバイスとして用い，街中で「ミクロ楽団」と呼ばれる仮想の生物を探すというアート・イベントである．2003 年に東京の渋谷を舞台として行なわれた．街中に置かれているパイロンの形をしたオブジェクトに PDA を近付けると，そこに隠れている「ミクロ楽団」をつかまえることができる．また参加者の PDA 間でコミュニケーションをとることもできる．最終的に全ての参加者がつかまえた「ミクロ楽団」を集め，一つの楽曲を作る．「ミクロ楽団」が生み出す歌の歌唱合成には，ヤマハ「ボーカロイド」[15] が使われている．このイベントは，総務省「企業 IT 化支援情報通信プラットフォーム構築に関する調査研究」の一環として実施された．

5. 音声インタフェースについての考察

音声インタフェースについての事例を元に，現状を考察する．

- 「Messa di Voce」「RE:MARK」「Hidden Worlds」は，音声という目に見えないものを視覚的に表現することに成功した作品として興味深い．特に「Messa di Voce」は非言語的な音声入力を活用したアート作品として他に類を見ない作品である．PitchPaint では音高の変化を用いて線を操作し，絵を描くことを可能とし，またその過程そのものを魅力的に表現している．
- 「n-Cha(n)t」と「FMS+ThirdEar」は音声認識を使用しているが，その結果としての単語を何か意味のある言葉として使用しているわけではない．結果としての単語を元に，詩的な言葉であったり，様々な画像や言葉の集積を表現するために用いており，結果としてランダムな言葉を生み出すための装置として機能している．
- 大半のゲームは，音声処理エンジンとしてミドルウェアを使用している．アート作品では，フリーソフトウェアなどを活用して制作されているが，まだ作品制作のための十分な性能を備えているとはいいがたい．「D SYSTEM」では，当初は音声認識を使うことを考えていたが，適切なライブラリが無いため断念している．アート制作の現場で利用しやすいライブラリの開発が望まれる．

6. おわりに

本稿では，音声認識，音響処理，歌唱合成を用いたアート・エンターテインメント分野の作品について，現状をまとめた．音声処理技術の成熟に伴い，様々な分野における応用事例が出てきていることがわかった．

謝辞 増井俊之氏，高林哲氏，山口優氏，坂井れいしう氏に感謝の意を表します．

参考文献

1) Golan Levin and Zachary Lieberman: “In-situ speech visualization in real-time interactive installation and performance”, *NPAR ’04*, pp. 7–14 (2004).
2) Golan Levin, Zachary Lieberman: “Messa di Voce”, 2003. http://tmema.org/messa/
3) Golan Levin, Zachary Lieberman: “RE:MARK” and “Hidden Worlds of Noise and Voice”, 2002. http://www.flong.com/remark/
4) David Rokeby: “n-Cha(n)t”, 2001. http://homepage.mac.com/davidrokeby/nchant.html
5) 師井聡子，笹田晋司，柴田良二「およぐことば」，2002.
6) ディレクション:小阪淳，プログラム:島田卓也「フキダマリ」，2000.
7) exonemo「FMS (FragMental Storm)」，2000. http://www.exonemo.com/FMS/indexJ.html
8) PACJAP，http://www.lafriche.org/pacjap/
9) 三輪真弘，“出品作品”SendMail”について — ある作曲家が体験したピッチ検出の実際と限界 —”，信学技報，SP96-115，pp.13–14，Feb. 1997.
10) 三輪眞弘「サクソフォン，ピアノとコンピューターのための SendMail」，1995. http://www.iamas.ac.jp/ mmiwa/BoysLabelMM.html
11) スキャンソフト，http://www.scansoft.co.jp/
12) ソニー:「S-FORCE」. http://www.sony.co.jp/Products/S-FORCE/
13) NTT アイティ:「Hipervoice」. http://www.ntt-it.co.jp/goods/vcj/voice/hipervoice.html
14) アニモ:「FineSpeech」. http://www.animo.co.jp/product/fs/
15) ヤマハ:「Vocaloid」. http://www.vocaloid.com/