SmartVoice: 言語の壁を超えたプレゼンテーションサポーティングシステム

李 翔　　暦本 純一*

概要.　国際的イベントでは, 共通言語での講演が求められることが多い. しかし, 語学力を短期間に向上させることが難しく, 共通言語が母国語ではない講演者にとって負担になる. また, 同時通訳などの方法は, コストが高い上, 講演者本人の発表を聞いているわけではないため, プレゼンテーションの自然さを損なってしまう. 本論文では, 言語の壁を越えてプレゼンテーションを可能にするシステム「SmartVoice」を提案する. SmartVoice は, 音声データ化した原稿を講演者の口の動きにあわせてスマートに送り出し, 講演者が自分のタイミングで講演できるようにする. 音声のイントネーションや強弱も, 口の位置と形状に基づいて制御するため, まるで講演者本人が直接喋っているように見える. 評価実験を行ない, 本人のプレゼンテーションと SmartVoice によるプレゼンテーションがほとんど区別できないことがわかった. SmartVoice のアプリケーションは, 多言語を用いるプレゼンテーションに限らず, 映像に音声・セリフを入れるアフターレコーディングにも適用可能である. 本論文では, 他分野における SmartVoice の応用についても議論する.

1　はじめに

国際学会などのイベントでは, 共通言語, たとえば英語で講演することが求められる. しかし, 共通言語が母国語ではない人にとって, それは大きな障壁となる. 仮に原稿を書いて読み上げることにしたとしても, 正しい発音で流暢に聞こえるためには多くの努力を要する.

言語が異なる講演のために, 同時通訳が利用されることがある. しかし同時通訳者を雇うコストや講演者とは異なるタイミングで音声が流れる不自然さなどの問題がある. また講演者の感情や非言語情報を伝達することが困難である. 予めプレゼンテーション原稿の音読を録音し, プレゼンテーションの時にそれを再生して口パクする方法もあり得るが, 講演者は録音のペースに合わせる講演をすることになり, 講演の自由さが大きく制約されてしまう.

ボイスシンセサイザーを用いることで, 原稿のテキストファイルさえあれば, コンピュータがどの言語の単語でも正確な発音で読み上げてくれる. しかし, 文章を朗読する時のイントネーション, 抑揚が単調であり, 聞き疲れするため, 人間の声の代用とすることはできない. 例えば通常の講演では, 大事なところで間を開けて聴衆の注意をひきつける, などが行われるが, ボイスシンセサイザーではこのような演出をすることは難しい.

本論文では, これらの問題を解決する手段として, 「スマートな口パク」に基づくシステムを提案する. スマートな口パクとは, まず原稿をナレーターに読んでもらうか, ボイスシンセサイザーを利用するかして音声データにしておき, それを講演者の口の動きに合わせて送り出す方式である. 講演者は自分のタイミングでプレゼンテーションすることができ, 大事なところで間を空けるなど, 語速を自由に制御できるようになる. また, イントネーションや音の強弱などを, 口の位置や形状から制御するようにしている. この結果, 聴衆からすると, 講演者本人が直接喋っているように見え, 同時通訳のような不自然さがない. この方式での講演は, 共通言語でのプレゼンテーションに留まらず, 聴衆に合わせた言語での講演 (たとえば中国で中国語で講演する) にも利用出来る. この, スマートな口パクに基づくシステムを SmartVoice と呼ぶことにする.

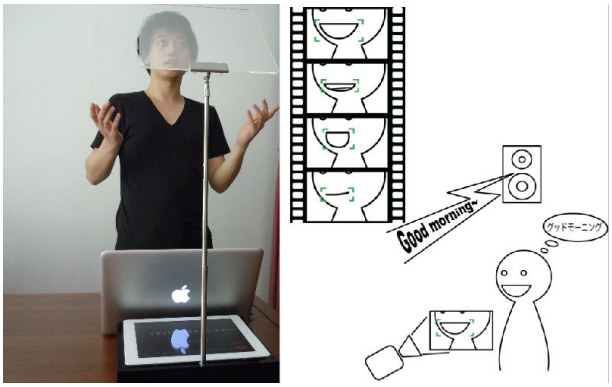

図 1. SmartVoice によるプレゼンテーション

2　SmartVoice

SmartVoice は, 再生制御およびピッチ, スピード, 音量のリアルタイム調整をプレゼンターの口の動きとその他表情にマッピングし, 音声が口パクのペース

* Xiang Li, 東京大学大学院 学際情報学府 Jun Rekimoto, 東京大学大学院 学際情報学府, ソニーコンピュータサイエンス研究所

に合わせて自然に流すシステムである．言語の壁を越えたスピーチやプレゼンテーションのサポーティングを始めとするさまざまなアプリケーションが考えられる．SmartVoice を用いたプレゼンテーションの様子を，図 1 に示す．

2.1 SmartVoice を用いたプレゼンテーション

実際に SmartVoice を用いたプレゼンテーションの仕方について説明する．

まず，講演原稿を用意し，それに基づいた音声データを準備する．音声データは，ユーザ自身もしくは自分以外のナレーターによって原稿を読み上げた録音か，ボイスシンセサイザーによって自動生成される音声を想定している．用意した原稿テキストファイル（.txt）および音声データをシステムに与えると，SmartVoice は半自動的に原稿と音声のフレーズレベルの対応付けを行う．そして音声と正しく対応付けられる原稿中のフレーズは，ハーフミラープロンプターに表示される．

利用者は，用意した音声データをシステムに与え，顔がカメラ領域に収められ，トラッキングできる状態で，プロンプター画面を見て，原稿中のフレーズを口パクすれば，そのフレーズと対応する音声が流れてくる．一つのフレーズを読み終えて口を閉じると，再生が止まり，プロンプターには次に読むべきフレーズが強調される．再び口パクをすると，今読むべきフレーズと対応する音声が流れるが，それまでに持続時間任意の一時停止ができる．また，イントネーションも前述したように，口パクのペース，口の位置と眉毛の位置によってリアルタイムでコントロールできる．

2.2 SmartVoice におけるスマートな口パク

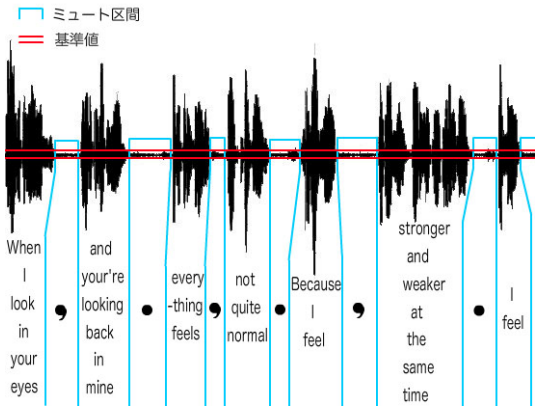

図 2. 音声と原稿からのフレーズ抽出と対応付け

2.2.1 音声の区切りおよび送り出し制御

実際に SmartVoice を使ってプレゼンテーションする時に，ユーザとインタラクションするためにプロンプターを利用している．ユーザがプレゼンター画面を見て，常に現在原稿のどこを読んでいるのかを確認することができる．そのために，音声とテキストは，正しく対応付けていることが必要である．SmartVoice は，この対応付けをシステムが半自動的に処理する．

通常の講演では，呼吸を入れることによって話のテンポを変えたり，一時停止することがある．なお，再生中の区切りに対し，違和感を感じさせないためには，最低限一つの単語の中での一時停止は避けるべきである．本システムでは，原稿テキストをカンマやピリオドなどのパンクチュエーションを区切りとしてフレーズを抽出し，音声データについては，フレーズは音声中に音響レベルが全音声データの平均値の一割以下で持続時間 0.1 秒以上の箇所（以下ミュート区間とする）によって分離された区間とする．一時停止は，ミュート区間内のみを許可する．ただし，特に録音の場合，ミュート区間中に，音響レベルが高いが，持続時間が極めて短いノイズは存在する．これらを見つけ出し，フレーズ抽出プロセスに無視されるよう，ノイズによって分断されるミュート区間が存在しないことを保証するためのアルゴリズムも用意した．(図 2)

原稿と音声の両方からフレーズが抽出されたら，本システムは後述するマッチングプロセスを用いて，原稿テキストと音声データの phrase-to-phrase の対応付けを行う．

プロンプターに表示される原稿テキストは，フレーズ単位で音声と対応し，ひとつのフレーズの音声を送り出したら，プロンプター画面には，自動的に次のフレーズのテキストに切り替わる．講演者はプロンプ他に表示されるフレーズを「口パク」で読み上げるか，概ねそのように口を動かすことで音声データが順次送り出されていく．

2.2.2 ピッチ，スピード，音量の制御

音声データは講演者の口の動きによって送り出されていくが，さらに，そのピッチ，スピード，音量も講演者の顔情報から制御できるようにした．例えば，ボイスシンセサイザー音声データを利用した場合でも，このようにして抑揚や感情表現を講演時に付与することができ，より自然な講演になる．

これらの制御は，人間が実際に喋っている時の原理を考えて，方針を決めた．具体的に，ピッチについて，頭を下げる時にピッチを下げ，逆の場合はピッチを上げるようにした (図 3 - b)．次にスピードについて，口の動きが激しければ激しいほど語速を高く設定する．最後に，音量について，目の開き具合によってコントロールしようとしたが，今回は眉毛の

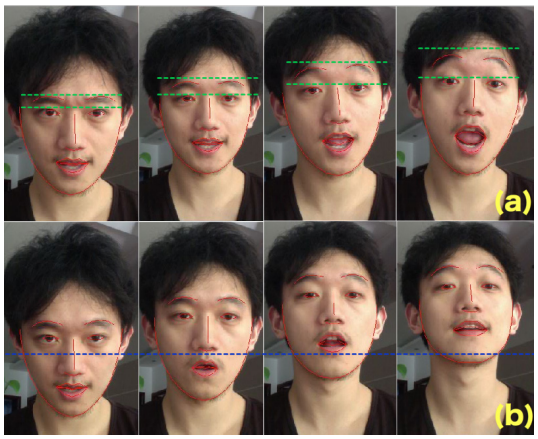

図 3. (a) 音量のコントロール (b) ピッチのコントロール

高さでコントロールすることにした（図 3 - a）. 理由は後述とする.

3 システム構成

本システムは, ホスト側とリモート側の 2 つの部分から構成されている. ホスト側は, ユーザの顔をトラッキングするためのカメラと, 音声を出すためのスピーカーが付いている諸処理を担当するノート PC からなり, リモート側は, プロンプターとして使われるタブレット PC とプロンプタースタンドからなる. なお, 両部分の間は, 無線 LAN 通信でコミュニケーションを取る.(図 4)

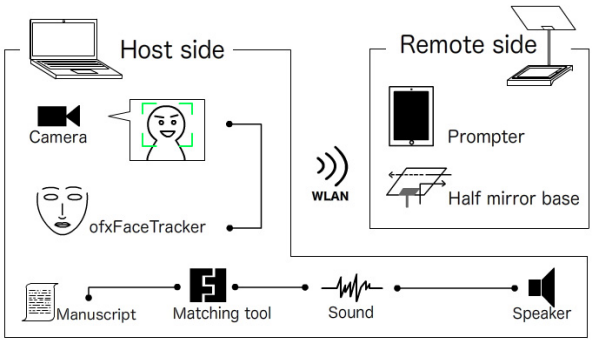

図 4. システム構成図

3.1 ホスト側

本システムは, プレゼンターに負担をかけず, より自然なパフォーマンスを実現するために, 使用中に全ての操作は, 口の動きや表情のみを使い, ハンズフリーに行う. プレゼンターは, 再生の一時停止と再開を決めることが可能のほか, 音声のスピード, ピッチそして音量をリアルタイムで変えることによって, ボイスシンセサイザーが生成した音声をより人間的にすることができる.

3.1.1 顔認識およびフェイストラッキング

SmartVoice は, 常時にプレゼンターの顔の位置, 口の動きおよび表情をトラッキングする. 本システムにおいて, openframeworks[1] 用の Add-on である ofxFaceTracker[2] を用いて, 人間の顔およびその傾きや大きさ, さらには目や鼻, 口, 眉といった顔のそれぞれのパーツの位置や大きさを立体的に検知してトラッキングする. 本システムのプロトタイプでは, Web カメラ付きノート PC を用いて, カメラから取得したリアルタイム映像からユーザの顔をトラッキングし, 画面にある顔および顔の各パーツの位置, 口の開閉パターンおよびそのペース, 眉毛の位置などを常時に記録し, 表示している.

3.1.2 Phrase-to-phrase の対応付け

前章で言及したように, 本システムにおいて, 原稿テキストと音声は, フレーズレベルで対応付けられている. この Phrase-to-phrase の対応付けの正しさを保証するために, 本システムの一部として半自動的マッチングツールを導入した.

マッチングツールはまず, 音声データ中のミュート区間を見つけ出し, フレーズを抽出し, 各フレーズの再生時間を配列にキャッシュする. 次に, 原稿テキストから抽出されたフレーズを, 順番にボイスシンセサイザーにかけて, 生成された一時的音声データの再生時間をそれぞれ取得して配列にキャッシュする.

ボイスシンセサイザーによって生成された音声の再生時間は, 語速の違いによって人間のナレーターによるものと著しく異なる場合があるため, マッチングツールは直接に前述の 2 つの再生時間の配列をマッチングするではなく, 音声データと原稿テキストのそれぞれ各フレーズの再生時間対総再生時間の比をマッチングの対象とする.

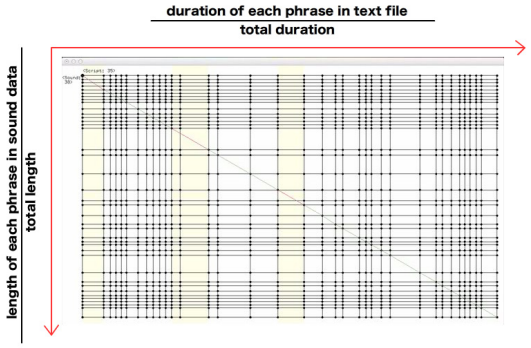

図 5. 自動マッチング結果を示すマッチングツールの画面

マッチングには DP マッチングアルゴリズムを利用している. 図 5 のように, 黄色に塗りつぶされた

部分は Miss match を表している．Miss match の原因は, 原稿と音声から抽出されたフレーズの総数が一致しないことだと考えられ, 主に 2 つのケースがある．一つは, 本来は区切りではないところで音声を止めてしまったため, ミュート区間とされた場合．もう一つは, ナレーターが原稿中の区切りを無視して続けて喋った場合である．いずれのケースに対しても, 本システムは手動修正をしやすくするための GUI をユーザに提供する．

3.1.3 再生コントロール, 感情表現付け

SmartVoice では, 一時停止の長さはユーザによって, 任意に決めることができる．ミュート区間中に一時停止するたびに, システムはそのミュート区間を丸ごと飛ばすことによって, ユーザが再び口を開くと, 音声がすぐ再生し, ディレイ無くプレゼンテーションを再開することが可能である．

前述の通り, ボイスシンセサイザーによるコンピュータボイスは, 自然なイントネーションに乏しく, 自然に喋っているように聞こえない．そこで, 本システムは, 再生時にリアルタイムで音声の音量変更, ピッチシフト, ピッチを変えずに再生スピードを変えることによって, プレゼンターの感情表現を付けることを可能にした．音声のダイナミックピッチシフトおよびストレッチは, C/C++ベースのマルチプラットフォームで動作するタイム・ピッチマニピュレーションライブラリ Dirac3 を用いて実現している．[3] ピッチシフトについては, 0.2 刻みでもとのピッチから上下 1 半音の変動幅に制限し, ストレッチについては, 0.7 倍速から 1.6 倍速の幅に制限している．音量について, 初期状態では 70 %にセットし, 再生時に 50 %から 100 %までの間で変化することができる．

3.1.4 マッピング

本システムは, プレゼンターをプレゼンテーションに専念させるために, すべての操作パラメータは, トラッキングされているプレゼンターの顔のパーツにマッピングしている．

再生の一時停止と再開は, ユーザの口の開閉にマッピングする．フレーズ再生後にプレゼンターが口を閉じ, 再び口を開くと, 次のフレーズの再生が開始される．また, 再生途中, つまりミュート区間以外の位置では, ユーザの口の開閉を問わず, 再生し続ける．これによって, 音声がスマートにプレゼンターの口の動きに合わせて流れてくるによって, まるで本人が喋っているように見えるという本システムのもっとも基本的な目的が実現される．

本システムでは, ピッチ, 音量およびスピードは, 再生時リアルタイムに変更を加えることができる．ピッチシフトについては, ユーザの顔位置にマッピングしている．具体的に, 再生時にユーザの顔の位置が初期状態顔の位置との縦方向の差で計算する．初期状態より顔を上げれば, ピッチが高くなり, 逆に, 頷くとピッチが低くなる（普通に喋る時に, 頷くと声帯が圧迫されるため）．音量の変更について, 前章で述べたように, 今回はユーザの目の開き具合にマッピングしている．しかし, 直接目の広き具合を取ると, 変化率が小さいため, 瞼と連動する眉毛の上下位置を採用した．ピッチシフトと似ており, 再生時にユーザの眉毛の位置が初期状態眉毛の位置との縦方向の差で計算し, 初期状態より眉毛を上げれば, 音量が高くなり, 逆の場合は音量が低くなる．スピード変更については, 再生時に口パクのペースにマッピングしている．具体的に, 今のフレームと直前のフレームにおいて口の面積の変化率を取って, 口の動きが激しければ, スピードを上げ, 逆にゆっくりと口を動かすと, スピードを落とす事が可能である．

3.2 プレゼンテーション時のインターフェース

3.2.1 プロンプター

本システムにおいて, プレゼンテーション時にユーザと実際にインタラクションをとるユーザインターフェースとして, 原稿を提示するためのプロンプターを設けている． なお, 本システムのプロトタイプ SmartVoice では, タブレット PC(iPad) をプ ロンプター画面とし, プロンプタースタンドとしてハーフミラーによるプロンプター台を利用している．ハーフミラー付きプロンプターを使うことによって, プレゼンターは講演会場を自然に見渡しながら原稿を読むことが可能となる．

3.2.2 ホスト側との協同作業

SmartVoice では, リモート側のプロンプターの稼働環境として iPad を採用し, リモート側とホスト側の間は, 無線 LAN 通信を用いてコミュニケーションを取っている．その理由は, なるべくプロンプターの機能をシンプルのままに保ち, プレゼンテーションできるまでの準備作業は, すべてホスト側で完結可能にしたいと考えているからである．そのために, プロンプターに表示すべき原稿は, ユーザが用意したテキストファイル (.txt) をホスト側プログラムで読み込み, 講演時に無線通信によりリモート側のプロンプター画面 (iPad) に伝送する．

実際にプレゼンテーションをする際に, ユーザとインタラクションをするのは, プロンプターであるため, ユーザは講演時にはプロンプター画面のみでシステムの挙動を確認できるようになっている．プロンプター画面は, 図 6 - c で示しているように, ホスト側でうまく顔を認識しているか否かを, ラベル「DETECTED!」でユーザが知ることができる (認識できてない場合は「???」と表示)．

音声と原稿テキストはフレーズレベルで対応付け

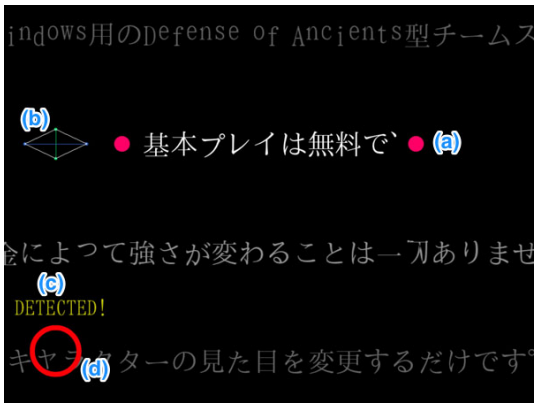

図 6. プロンプター画面

ているため, プロンプター画面にも, フレーズ単位で原稿をユーザに提示する. 具体的に, プレンゼンターが今読んでいるフレーズだけもっとも見えやすくし, 読み終わったフレーズはスライドアップするようにしている. (図 6 - a)

ピッチや音量の変化のようなすぐに分かる変化と異なり, 本システムを用いてナレーションの再生スピードを変える場合, 現在の再生速度の目安をユーザに提示する. プロンプター画面では, 口のイメージ (図 6 - d) が, 喋る速度 (ナレーションの再生速度) に合わせて, 開閉アニメーションの速度が変わる. 一時停止の時は閉じるままである. また前述のように, 本システムにおいてスピードを決めるのは口の面積の変化率であるため, そのメタファとして, 図 6 - b で示す 2 つのバーをプロンプター画面に追加した. この 2 つのバーは, ダイナミックに口の高さと広さを表し, 今取られているパラメータをユーザに提示する.

3.3 性能評価

本システムにおいて, フレーズレベルの音声と原稿テキストとの対応付けは, ホスト側が半自動的にしてくれる. ボイスシンセサイザーが生成するナレーションの場合, パンクチュエーションが遵守されるため, 対応付けは正しく行われる. ところが, 人声によるナレーションの場合, 時々パンクチュエーションを無視したり, フレーズの途中で呼吸を入れたりするナレーションがあるため, 対応付けが乱れることがある. その場合, 本システムがのマッチングツールが提供するインタフェースを用いて対応付けを手動修正することができる.

本システムに使われるノート PC に付いているカメラは, 最大解像度 720p で最大 30fps の撮影機能を備えている. 本システム稼働中, プレゼンターによるすべての操作に対し, ホスト側プログラムの FPS は 55fps を上回っており, ディレイはほとんど感じられない.

4 ユーザテスト・評価

SmartVoice の効果を確認するために, 利用者評価実験を行った. 通常のスピーチをしている動画と, SmartVoice を用いてスピーチをしている動画とをビデオ編集によって切り替えたものを実験参加者に見てもらい, どの部分が「口パク」(SmartVoice による箇所か) を判定してもらった. 判定結果はマウスボタンの押下により入力してもらった. 声色の変化による影響を避けるために, この実験では本人が読み上げた音声を, 本人が SmartVoice によって再生するということにした. 8 名 (コンピュータサイエンス関係学科の大学院生) が実験に参加した.

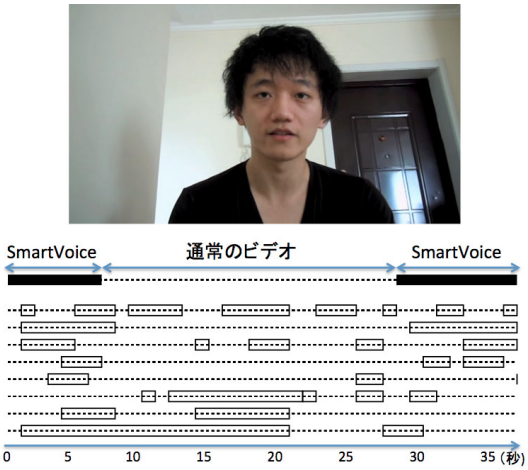

図 7. 生声原稿読み上げ v.s SmartVoice の実験結果

実験結果を図 7 に示す. それぞれの参加者が SmartVoice によるものと判定した部分を白抜きの矩型で示す. この図から明らかなように, SmartVoice ではない部分も口パクと判断した箇所がある一方で, SmartVoice による部分も本人の生のプレゼンテーションだと判断している箇所もある. この実験から, SmartVoice による音声対応は本人によるプレゼンテーションと比較しても自然であることがわかった.

5 議論

5.1 関連研究

フェイシャルアクションを使って, 何らかの制御にマッピングする前例として, Lyons らによる The Mouthesizer がある. The Mouthesizer では, ミニヘッドマウント CCD カメラを用いて, 口の中の影領域をトラッキングし, MIDI のエフェクト付きや楽器の操作などをユーザの口の動きにマッピングしている. [4] また, 口の高さ, 広さの他, 舌の動きと位置も使われている. Lyons らは, 楽器を演奏するミュージシャンに両手以外のコントローラーを与えた点では, The Mouthesizer はフットペダルと同じだが,

Mouth コントローラーはペダルより直感的で身に着けやすいことを証明し, フェイシャルアクションによるマシンインターフェイスの可能性も言及した. [5]SmartVoice はコンセプト上, The Mouthesizer のとシェアするが, 操作中に変わった表情などを取り入れていないため, プレゼンテーション中にユーザは自然な状態を保つことができる.

また, 同じくフェイストラッキングを用いるが, 音声ではなく, リアルタイムで CG キャラクターの表情を制御する研究もある. [6]

5.2 ボイスシンセサイザーによるビデオ作品の音声付け

従来, デモビデオなどにナレーションを入れたい場合, 特にナレーションが母国語ではない時に, 原稿のテキストファイルを用意し, ボイスシンセサイザーによって音声化する手法がある. しかし, 前述の通り, ボイスシンセサイザーは単語一つ一つをキレイに発音してくれるが, 抑揚に欠くため, 機械的に聞こえる難点がある. 本システムのように, 抑揚やタイミングを人間がコントロールすると, ボイスシンセサイザーによる声をビデオのナレーション作成やアニメーションの人声の作成などにも応用できると考えている.

5.3 アドリブの可能性

今まで述べてきたように, 本システムに使われる音声は, 事前に作っておかないといけない. しかし, プレゼンテーションの場合, 最初から最後まで原稿に準して喋る必要がなく, 途中で雑談を挟むことも考えられる. SmartVoice による音声と生声のスイッチによる違和感を隠蔽し, それを気づかせないようにすれば, SmartVoice と生声の間で自在に切り替え, プレゼンテーション中のアドリブも可能と考えている.

5.4 アフターレコーディングの可能性

ドラマや映画の作成は, 撮影と音声はそれぞれ単独に収録し, プロの声優のアフターレコーティングによってセリフを映像に入れるのは一般で的である. 本システムは, リアルタイムにカメラからキャプチャーした映像に限らず, 収録済みの映像中のキャラクターの顔をトラッキングして利用することもできるため, 出来上がった映像を見ながらアフターレコーディングを行わなくても, SmartVoice を用いれば, タイミング正しく映像に音声を入れることが可能である.

5.5 歌, 外国語練習への応用

繰り返しになるが, SmartVoice は「まるでその人が生で喋っているように見える」のようなシステムの実現が目的である. 広義的な「人間が喋る」という点から, 本システムが支援できるのは, スピーチやプレゼンテーションに限らないはずだ.

例えば, カラオケと類似するプロンプターを有するため, 練習しやすい上, SmartVoice はリアルタイムでピッチシフトとスピード変更もできるため, 音感・リズム感の悪い人に, ピッチ感覚とリズム感を身に着けさせ, 歌が上手くなることができるかもしれない.

また, 外国語学習者の中で, その言語をできるだけネイティブっぽく喋れることを望む人ことが多い. しかし, 発音やイントネーションを真似して, たくさん練習してそれを自分が喋る時の習慣とするのは結構大変である. 外国語学習者用に SmartVoice にネイティブ声との比較機能を付けることによって, ネイティブ風のイントネーションを身につけることができるかもしれない. さらにもっと細かな顔筋肉がトラッキングできれば, SmartVoice による発音矯正も不可能ではないと考えている.

6 結論とまとめ

本論文では, 言語の壁を超えたプレゼンテーションサポーティングシステム「SmartVoice」を提案した. 音声データを分析し, 原稿テキストとフレーズレベルの対応付けを実現することによって, ユーザの口の動きに合わせてスマートに音声を送り出すことを可能にした. また, ボイスシンセサイザーによる機械的なナレーションのイントネーションを, ユーザの口の形状・位置によってコントロールすることを実現した. 本システムの効果を検証するために, 評価実験を行い, SmartVoice による音声対応は本人によるプレゼンテーションと比較しても顕著な不自然がないことがわかった. 今後, 他の分野における SmartVoice の応用が期待できると考えている.

参考文献

[1] openframeworks. http://www.openframeworks.cc/

[2] ofxFaceTracker. http://github.com/kylemcdonald/ofxFaceTracker

[3] Dirac3L. http://dirac.dspdimension.com/

[4] L. Michael J., H. Michael, and T. Nobuji. The Mouthesizer: A Facial Gesture Musical Interface. In *Conference Abstracts, Siggraph 2001*, pp. 230, 2001

[5] L. Michael J., T. Nobuji. Facing the Music: A Facial Action Controlled Musical Interface. In *Proceedings, CHI 2001, Conference on Human Factors in Computing Systems*, pp. 309-310, 2001

[6] W. Thibaut, B. Sofien, L. Hao and P. Mark. Realtime Performance-Based Facial Animation. In *ACM SIGGRAPH 2011 Papers, SIGGRAPH ‘11*, pp. 77:1-77:10, 2011