Vol.2016-CE-133 No.11
2016/2/13

英語プレゼンテーションに関する学習を支援するソフトウェアの開発

後藤健太[†1]　西原悠貴[†1]　中平有樹[†1]
吉本定伸[†1]　小嶋徹也[†1]　堀智子[†1]　鈴木幸一[†2]

概要：社会のグローバル化が進み，国際的な会議や大会において英語を用いたプレゼンテーションに関するスキルの重要性が高まっている．そこで学習者のうち特に学生が，効果的な英語プレゼンテーションを行うことができるように学習を行うことが必要である．そのような学習を行うため，リアルタイムで自分の発している音声とお手本の音声を視覚的に比較できるシステムを開発する．

キーワード：英語プレゼンテーション，学習支援ソフトウェア，ユーザインターフェイス，音声解析，声質変換

Development of English Pronunciation Software for Oral Presentations

KENTA GOTO[†1]　HARUKI NISHIHARA[†1]　YUKI NAKAHIRA[†1]
SADANOBU YOSHIMOTO[†1]　TETSUYA KOJIMA[†1]　TOMOKO HORI[†1]
KOUICHI SUZUKI[†2]

Abstract: As globalization proceeds, the importance of presentation skills in English at international conferences have been widely recognized. Japanese students, in particular, are expected to improve English proficiency and presentation skills. In this study, we developed a computer-based learning system which assists learners to practice English pronunciation and prosody. This system enables learners to compare their pronunciation and model sounds visually in real time.

Keywords: Presentations in English，Learning Software，User interface，Acoustic Analysis of English Sounds, Voice Quality Conversion

1. はじめに

英語でのプレゼンテーションの際には発音の強弱や抑揚の変化，単語と単語の間にポーズを設けるなどの手段によって伝えたい事柄を強調する．しかし，既存の英語音声学習ソフトウェアは難解な波形表示によってお手本の音声と学習者の音声を比較するため，音声学の知識がない学習者が学習を行うのは困難である．そこで，お手本の音声と学習者の音声を直感的に理解できる形式で比較できるようにすることで，音声学の知識がない人でも学習を行うことのできるソフトウェアを開発する．

昨年度までの研究において，音声特徴を抽出し，プレゼンテーション時と音読時のピッチの変動幅の比較を行い，音声学習ソフトウェアの機能検討及びソフトウェアのプロトタイプを作成した[1][2].

本稿ではこれらの成果を踏まえ，声質変換技術に基づくお手本音声変換に関する検討と実験及び音声の特徴解析と比較，ソフトウェアにおける音声特徴表示の改善と英文表示の検討と実装を行う．

†1 東京工業高等専門学校
National Institute of Technology, Tokyo College.
†2 鈴木幸一事務所
Kouichi Suzuki office

2. ソフトウェア設計の検討

2.1 ソフトウェアの機能の検討

昨年度，英語プレゼンテーション学習に必要な機能の検討等が行われた．それを以下に示す．

① お手本音声の読み込み，及び再生
② お手本音声に対するピッチ変更
③ マイクを用いた学習者の音声の取得
④ 音声のピッチ抽出および画面への表示
⑤ 学習者の音声とお手本音声のピッチの比較

今年度はこれを元に，開発するソフトウェアの処理の流れを検討した．(図 1)

まず，お手本話者の音声に対してユーザーが真似しやすい音声への変換を行う．その音声を出力するとともに特徴分析を行い，結果を画面に出力する．

次に学習者はお手本音声に対する出力を確認しながら音声を入力，その音声の解析を行いリアルタイムに画面出力することによってお手本音声との比較を行う．

2.2 ソフトウェア構築の検討

ソフトウェアの構築には昨年度と同様に.NET C# WPF を用いることとした．音声の取得と出力に関しては DirectX の Direct Sound を用いた．

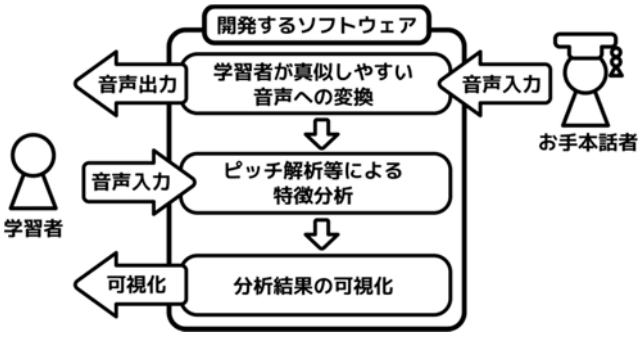

図 1　ソフトウェアの構成

3. 声質変換についての検討と実験

学習者がより効率よく学習を行うため，学習者の声質に似た音声で学習を行うことが効果的であると考え，お手本音声の声質変換について検討と実験を行った．

3.1 声質変換

声質変換とは，ある話者の音声の声質を別の話者の声質へ変換を行うことである．音声から声質を決定するパラメーターを抽出し，変換したい音声へ適用することによって音声を変換する．今回はパラメーターとして声帯振動による音源の情報である基本周波数(F0)と，声道の形状情報であるスペクトル包絡の 2 つを用いて声質変換を行う．

本研究では声質変換に用いるパラメーターを抽出するために，既存の音声分析合成システムである WORLD[3]を用いた．WORLD では音声を基本周波数，スペクトル包絡，非周期性指標の 3 つのパラメーターに分解することができる． これを用いた声質変換の流れを図 2 に示す．

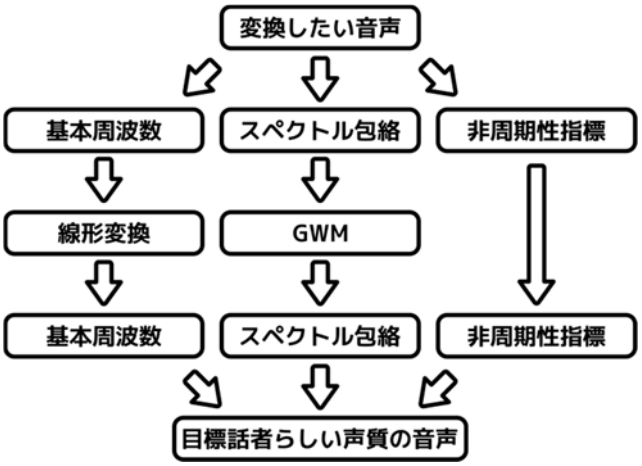

図 2　声質変換の流れ

3.2 基本周波数の変換

学習者に対応したお手本音声を生成するため，まずは声の高さを変換することを考える．声の高さを変換するために変換元の音声から F0 系列を抽出し，平均と標準偏差を考慮して

$$P_t^{(y)} = \frac{P_t^{(x)} - \mu^{(x)}}{\sigma^{(x)}} \cdot \sigma^{(y)} + \mu^{(y)} \quad (1)$$

により変換を行う．ここで，$P_t^{(x)}$は変換元音声の F0 系列であり$\mu^{(x)}$,$\sigma^{(x)}$はそれぞれ$P_t^{(x)}$の平均と標準偏差である．また，$\mu^{(y)}$,$\sigma^{(y)}$は目標話者の平均と標準偏差を示す．$\mu^{(y)}$,$\sigma^{(y)}$については蓄積された目標話者の音声を用いて計算する．

3.3 スペクトル包絡の変換

現在までに，声質変換の研究では数々の統計的スペクトル変換法が提案されている．本研究では代表的な方法の一つである混合正規分布モデル(GMM: Gaussian Mixture Model)を用いたもの[4]を採用する．

GMM でモデリングする際には，スペクトル包絡をメルケプストラム係数に変換して特徴量として用いる．メルケプストラム係数とは，人の周波数知覚特性を考慮して重み付けをした特徴量であり，任意の次元数で表現することができる．

この変換には 2 つの実現手法があり，1 つはフレームごとの変換，もう 1 つは系列全体での変換である．

3.4 実験

基本周波数の変換及びスペクトル包絡変換の 2 つの手法についてそれぞれ変換を行った．開発しているソフトウェアで動作させるという目的を達成するため，以下の 4 つを評価基準とした．

① 目標話者に声質が近づいているか.
② 変換元音声の言語内容が保たれているか
③ 十分な音質であるか.
④ GMM の学習にどの程度の時間を要するか.

これらについてスペクトル包絡変換の 2 つの手法で得られた変換結果の主観的評価を以下に述べる．

①に関しては両手法で目標話者に声質が近づいていることが確認できた．一方で，人間らしい声質が失われ，ロボットのような不自然な音声になってしまっていると感じられた．声質の変換精度は両手法に大きな差は感じられなかった．

②に関しては変換した音声のほとんどが変換元音声の言語内容を保っていたが，一部異なる音に変換されてしまったものがあった．両手法で同様の現象が生じていた．具体的には，curious，ask，her，him の単語で，子音部分がまったく別の音に変換されたり，消えてしまったりする現象が観測された．

③に関しては，両手法とも変換元音声に比べ大きく音質が劣化していた．特に前者の手法による変換音声の方が，後者の手法に比べノイズが多く,劣化の程度が大きかった．

④に関しては，前者の手法では約 20 分，後者の手法では約 90 分，学習に時間を要した．ここでは変換に要する時間は学習にかかる時間に比べ十分小さいので考慮していない．

4. 音声の特徴解析

学習者及びお手本音声の特徴を解析することによってその特徴を視覚的に確認，比較することができる．そこで音声の特徴解析について検討，音声の比較を行った．

4.1 ピッチの抽出

音声波形を短い時間で区切り，自己相関関数に通すことでその時間でのピッチを抽出する．自己相関関数の計算には FFT を用いる．

4.2 音響インテンシティ

ある音の大きさを基準値と比較し，常用対数によって表現したものを音響インテンシティ，あるいは騒音レベルという。振幅値から音量を解析するために，この数値を用いて，音量の確認を行う。変換は式(2)によって計算される。*f(x)*は音響インテンシティ，変数 x は 0 以上の数，a_nは振幅値を与える。

$$f(x) = 20\log_{10}\sqrt{\frac{\sum_{n=x}^{1023+x} {a_n}^2}{1024}} \quad (2)$$

音声中，音響インテンシティの値が低い箇所では発音がなされていない可能性が高く，高周波ノイズによりピッチ抽出がうまくいかない場合がある．よってこういった箇所ではピッチ抽出を行わない事によって波形を表示した際のノイズを消すことができる．

ピッチ抽出と音響インテンシティを用いて表示した波形を図 3 に示す．

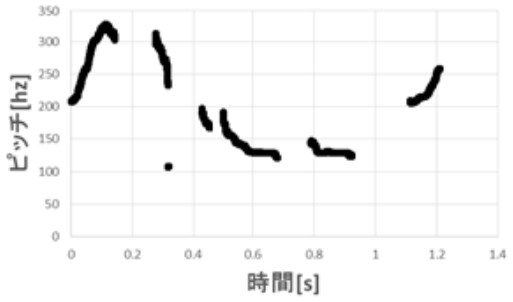

図 3　ピッチ抽出

4.3 ピッチの比較

音声の特徴分析結果を利用し，お手本と学習者のスピーチをピッチの比較による採点を行う。基準となるお手本のスピーチのピッチに，学習者がどれだけ沿った発音をしているかを確認する。しかし，人によって声の高さは様々であり，抽出したピッチをそのまま比較することはできない。ピッチの変化の様子は一致するが，ピッチの高さが異なる場合，採点・評価をすることは難しい。これを解決するために，セミトーンによる比較を行う。セミトーンは，人による声の高さを比較する際に用いられる指標である。変換は式(3)によって計算される。*Semitone* はセミトーン，f_1は変換する周波数，f_2は基準となる周波数が入る。ここでは，f_1に抽出したピッチ，f_2に平均のピッチを与える。図 3 のピッチをセミトーン変換したものを図 4 に示す。

$$Semitone = 12\frac{\log_{10}\frac{f_1}{f_2}}{\log_{10} 2} \quad (3)$$

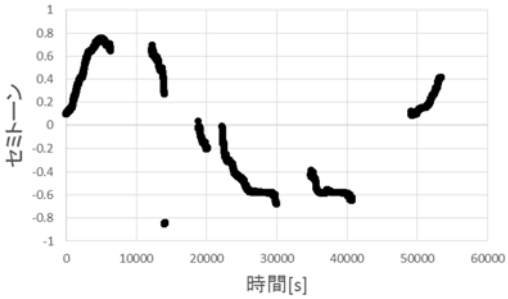

図 4　セミトーン

4.4 作成したプログラムの評価

作成したプログラムについては，フリーソフトウェア Praat[5]を用いて，評価を行う。図 5 に Praat を用いて解析を行ったものを示す。図 3,4 に示したプログラムによるピッチの変化とほぼ同様な結果が得られている。ピッチが表示されていない部分は，音響インテンシティの値が低いため，表示されていない。その部分も似たように表示されているので，音響インテンシティによる判定も行えている。

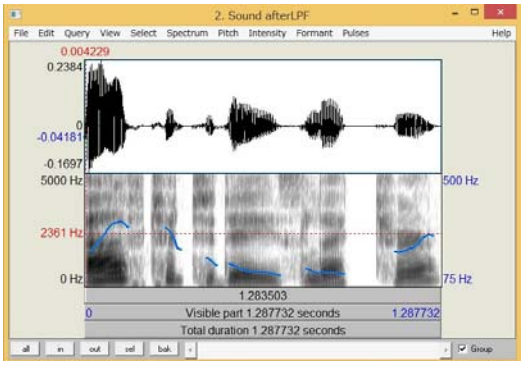

図 5　Praat による音声解析結果

5. 入力音声を表示するシステムの検討と実装

従来のソフトウェアに用いられている音声波形表示ではお手本音声及び学習者の音声の特徴を視覚的に捉えづらい．そこで，音声特徴の表示を改善することによって学習の効率化を図る．

5.1 音声特徴表示の検討

昨年度は，図 6 のように音声特徴の表示にピアノロール方式が採用された．しかし，この方法ではお手本になる音声の波形との差異は直感的に理解することができないと考えた．そこで，お手本の音声はピアノロールで表示し，学習者の音声をピッチ波形で表示する方式を考案した．しかしながら，この方法での表示も，音声の強弱を表示できないというデメリットが存在する．そこで，図 7 のようにピッチ波形に太さ・色の変化を付加し，音の強さと高さを表現する方法を考案した．この方法ならば，学習者が自分の音声の大きさやピッチを直感的に把握することができると

考えられる．

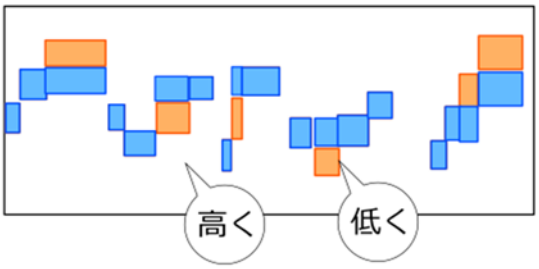

図 6　音声特徴のピアノロール表示

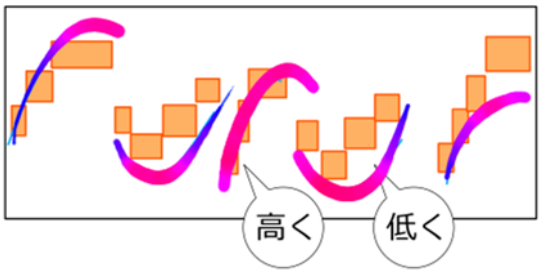

図 7　音声特徴のピッチ・音量波形表示

5.2 入力音声を表示するシステムの実装

学習者が自分の音声を目で確認しながら発音を行うことで学習効果がより高まると考え，前章で検討したようなインターフェイスをリアルタイムで表示するシステムの開発を行った．これは学習者の音声を取得し，描画する円の半径と縦座標を設定，それを連続的に表示することで図 8 のような線で表示するシステムである．

本表示システムにおいては鈴木氏からも「音声の特徴を表現できている」との良い評価を得られた．

図 8　太さと高さの変化する線の表示

6. 英文表示システムの検討

より視覚的に学習を行うためには，流れている音声に合わせて英文やコメントを表示することが有効であると考えられる。そこで，音声に合わせて英文を登録，再生が行えるシステムを開発した．その外観を図 9 に示す．

まず，英文の保存・再生に必要なファイルフォーマットを検討した．その結果，汎用性，可読性を考慮しデータ形式として広く用いられていてテキストファイルで構築できる Json 形式を用いることとした．

開発したシステムの操作方法を以下に示す．

まず英文を表示させたい音声ファイルを開き，表示させたい英文をテキストボックスに入力する．その後，音声を再生しながら単語の切れ目でボタンをクリックすることにより単語ごとに表示する秒数を記録する．記録されたファイルを読み込むことで，先ほど登録したタイミング通りに英文が流れて表示される．

実際に使用したところ，手打ちでタイミングを登録するため正確ではあるが，非常に手間がかかるものとなった．今後は，音声の切れ目などを自動で検出しタイミングを登録できるよう改良する必要があると思われる．

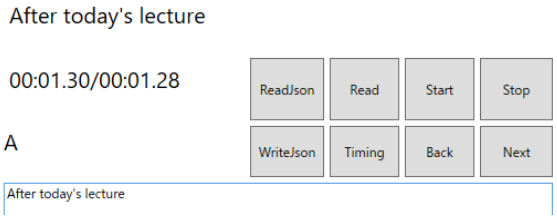

図 9　英文登録・表示システム

7. まとめ

ソフトウェア設計の検討とインターフェイスの検討，既存のモデリング手法を利用してモデル音声の声質を学習者のものに近づけるための検討と実験，音声の特徴解析，そして実際にマイクから入力した音声をリアルタイムに表示するシステムの開発，英文表示システムの検討と実装を行った．

今後は客観的な指標を用いて声質の変換精度を評価することや，変換によって変換元音声の重要な特徴が失われないような方法の検討，学習者音声の得点化，音声表示に関して動作の安定性を図るとともに各システムの統合が必要になると考えられる．

謝辞　本研究を進めるにあたり，アドバイスをいただいたコンテンツデザイナーの鈴木幸一氏に感謝の意を表します．本研究は，JSPS 科研費 25370680 の助成を受けたものです．

参考文献

1) 今井美和花， 松永竜太郎， 小嶋徹也， 吉本定伸， 堀智子， 野口ジュディー・津多江， “音声信号処理に基づく英語プレゼンテーション音声の特徴分析”， 電子情報通信学会技術研究報告，vol.113， no.415， pp.5-10 ， (2014).

2) 齋藤光， 橋積裕紀， 吉本定伸[他]， “学生を対象とした英語プレゼンテーション学習支援ソフトウェアの開発”， 教育システム情報学会研究報告， vol.29， no.5， pp.119-122 ， (2015).

3) 森勢将雅，音声合成システム WORLD，http://ml.cs.yamanashi.ac.jp/world/introductions.html

4) Tomoki Toda*et al.*，” Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory”， Proc. of IEEE2007，Glasgow，Scotland， pp. 2222 - 2235 (2007).

5) Paul Boersma, David Weenink , Praat http://www.fon.hum.uva.nl/praat/