音声認識システムを用いた発話訓練システム構築の試み*

上野 歩美*1　中川 彰*2　苣木 禎史*1　宇佐川 毅*1

*1(熊本大学工学部) *2 (熊本大学大学院自然科学研究科)

1. はじめに

社会の国際化や1990年の入管法（出入国管理及び難民認定法）の改正に伴い，外国から日本への入国者は年々増加する傾向にある。平成14年度末の法務省の統計によると，日本の総人口のうち約1.45%を外国人が占めていることになる[1]。また日本語学習者数も増加しており，国内で約9500人（平成12年文化庁調査），海外で約210万人（平成10年，国際交流基金調査）の外国人が日本語を学習している[2]。このように，様々な国からの外国人の流入をうけ，日本語学習者の多国籍化が進み学習目的が多様化したことで，様々な種類の日本語教材の必要性が問われている。

本研究では，日本語初習者に対する日本語学習の支援を目的とした発話訓練システムの構築，また，データベース上にWEB教材として使用するための日本語学習者用テキストの作成を行っている。さらに，日本語特有の発話現象である長母音と母音の無声化の知覚訓練も目的として挙げている。

本報告では，システムに以下のような機能を持たせることを提案する。

- 教示用の音声，日本語テキストを提示
- 学習者の発話を認識させる
- 学習者と対話を行う擬人化インタフェースを作成する
- 個人情報をデータベースに記録する
- クラスタシステム上で動作させる
- カメラ等で学習者の口唇の動作を記録・認識させる

学習者に対し，録音された音声や日本語テキストを提示することで発話訓練を行なう。学習者の発話を認識させるため，汎用大語彙連続音声認識エンジン Julius[3](以下 Julius) を用いる。また擬人化音声対話エージェント Galatea[4](以下Galatea) を用いて教師らしいインタフェースを作成することで，学習者が学習意欲を持続できる環境を提供することを考えている。また，学習者の進捗状況や癖などの学習の情報を基にして，それぞれの学習者個人に合わせた学習法を提案するためデータベースを作成する。提案するシステムはクラスタシステム上で動作させる。クラスタシステムをサーバ，学習者のPCをクライアントとし，ネットワークを経由して使用することによりインターネットを介した利用が可能である。つまり，教材を日本で作成し，直ちに国外で利用できるという物理的な距離と時間の制約を取り除くことが可能である。また，カメラ等を用いた画像認識システムに発話時の口唇の動作を認識させ，学習者に正しい口の動作を認知させる事も考えている。

2. 発話訓練システムの概要

現在構築している発話訓練システムのブロック図を図1に示す。発話訓練システムは，システムの基本部分を入力部，出力部，音声認識部，総合統括部の4つで構成しており，対話型システムを音声合成部，画像認識部の2つで構成している。さらに，処理能力の確保のためクラスタシステム上で動作させる。また，入力部は音声入力部，画像入力部の2つから，出力部はテキスト出力部，音声出力部，画像出力部の3つからなる。統合統括部にはWEB教材を作成するためデータベースを，音声合成部にはGalateaを用いる。

現在完成しているのはクラスタシステム・音声入力部・テキスト出力部の3つである。画像入力部・音声出力部・画像出力部・音声認識部・総合統括部・対話型システムは来年の春までを目標に構築中である。

---

* A configuration of an uttrance training system based on a speech recognition system

By Ayumi UENO, Akira NAKAGAWA, Yoshifumi CHISAKI, and Tsuyoshi USAGAWA
(Kumamoto University)

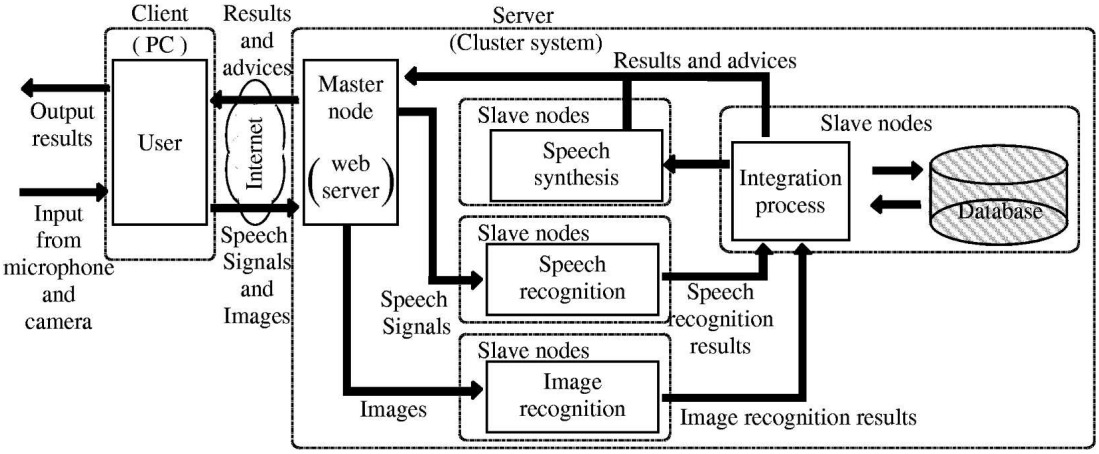

図 1．発話訓練システムの構成

今回提案する発話訓練システムは外国人の日本語初習者を対象としている。本システムでは学習者に文章や画像を提示し，それに対して選択肢の選択や音声による応答をしてもらうことで学習を進める。学習者の選択や応答の結果からその先の学習内容を変化させる。これは一度行なった学習でも単調な作業のくり返しにしないことで，学習意欲を保つことを目的にしている。つまり，学習者が主体となって自主的に学習を行うことが出来るよう，単純な発話練習だけでなくシステムとの対話を行う事が出来る教材を作成する。さらに，教示用の口唇の動きを画像として送信し，カメラ等で記録した学習者の口唇の動きと比較する事の出来る機能を持たせる。

2.1　ハードウェアシステム構成

本研究では図 2 に示す構成のクラスタシステムを構築している。高い処理能力を実現するため，発話訓練システムのサーバをクラスタシステム上で動作させる。クラスタシステムとは，複数のサーバを単一のシステムとして協調動作させることで，高い可用性と拡張性を提供するシステムのことである。

クラスタシステムにおいて通信を行う独立体を node と呼び，クラスタシステムは，全体の管理を行なう master node と，演算を行なう slave node からなる。master node は学習者からのデータを slave node に渡す。データを受信した slave node は演算を行ない，結果を master node に返す。master node は返ってきた結果を学習者に返す。

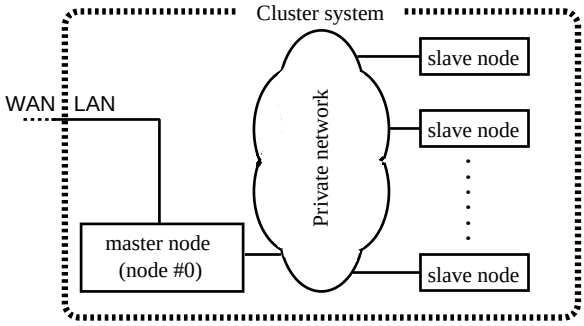

図 2．クラスタシステムを用いた実時間音響信号処理システムの構成

2.2　システムの基本構成

本システムでは入力部にて取得した学習者の発話を音声認識部で認識し，総合統括部で解析した結果やアドバイスなどを出力部で学習者に提示する。以下にシステムを構成するブロックについて説明する。

2.2.1　入力部

入力部は音声入力部と画像入力部からなる。画像入力部は現在構成中である。音声入力部では学習者のクライアント PC に対し，以下のことを要求している。

- A/D 変換機能を備える
- マイクロホンによる入力が可能である
- 高速なインターネットへの接続環境
- Java スクリプトに対応したブラウザが動作する環境

発話訓練システムに対する入力は，クライアント側からインターネットを介して送信されるマイクロホンで観測された音声信号である。観測する際のサンプリングレートと量

子化精度は音声認識部の制約からそれぞれ16kHzと16bitとなる。よって想定しているシステムではおよそ250kbpsが必要となり，そのようなネットワーク環境での利用を考えている。入力部はJavaスクリプト等のプログラミング言語を使用し，クライアント側からデータを受信する。

2.2.2 出力部

今回，文字・音声・画像などのデータを双方向に通信するためJavaアプレット等を用いた環境を作成する。音声・画像出力部は現在構成中である。テキスト出力部はJavaアプレットによりブラウザ上に画像や文字を表示させ，学習者にたいし学習用のテキストや発話の分析結果等を提示する。Javaアプレットはサーバから送られた指示に従い，ユーザに情報を提示する。

2.2.3 音声認識部

音声認識部の音声認識エンジンとして一般に無償で公開されているJuliusを用いる。Juliusは言語モデルとして単語N-gramを用い，音響モデルとしてHMM(Hidden Markov Model)を用いる。認識率は，20000語彙の読み上げ音声で90%以上である。汎用性も高く，発音辞書や言語モデル，音響モデルなどの音声認識の各モジュールを組み替えることが可能であり，単語認識，連続単語認識も可能である。また，プログラムのソースが公開されているので改変して使用することができる。

音声認識部は現在構築中であり，現在言語モデルには新聞記事より学習したものを用いているが，口語に対応したモデルに変更することを考えている。

2.2.4 総合統括部

統合処理部は音声認識部と画像認識部から送られてきた情報と，データベース内のWEB教材の情報により学習者に送信する学習用データを決定する。また，教材データベースとしてPostgreSQLを用いる。

PostgreSQLはほとんどのUNIXおよびUNIX互換プラットフォームで稼働する本格的なデータベースである。特徴としてユーザー定義型の配列の作成が可能であり，C言語やJava言語などのプログラミング言語をサポートしていることが挙げられる。さらに，サーバ・クライアント方式によりデータベースを利用するクライアントと，データベースエンジンを提供するサーバが完全に独立しているため，クライアントが異なるプラットフォームでも問題無く動作する。また，日本・韓国・中国のEUC，Unicodeなどに対応しており，サーバ側とクライアント側で異なる文字コードが使用可能である国際化対応のデータベースなので，外国人向けの教材としてのデータベースも作成しやすいと考えられる。

構築中のデータベースには日本語テキストと学習者の学習履歴や進捗状況などの個人情報を収める予定である。

2.3 対話型システムの構成

発話訓練を行うにあたり，教示用の発音を聞かせ復唱させるだけでは学習の効率が悪く，文のイントネーションは改善されやすいが語アクセントは改善されにくいという傾向が挙げられている[5]。つまり学習者に妥当な発音基準を形成させるには，単にモデル音声を聞かせるだけでなくなんらかの韻律規則に関する情報を与えたり，視覚的な補助を何か与えたりする必要があると述べられている[5]。そこで，視覚的な補助として口唇の動作を与えることを考える。正しい発声を行うには正しい口唇の動作が必要である。よって，使用者の口唇の動きを画像認識技術を用いて自分の口唇の動きと教示用の口唇の動きの違いを認識させることにより，使用者に適切な口唇の動作の指示する。

音声合成の音声は不自然な発音であるため，教示用音声には主に録音した音声を使用する。補助的な部分に音声合成を用いる。そこで本システムでは音声と音声合成技術･画像認識技術を使い，対話型システムを構築する。録音音声による教示用音声と，教示用の口唇の動作を記録した動画を送信する。また，教示用画像に合わせて学習者が発声する際，その口唇の動きを画像認識技術を用い認識させることにより，より高い効果で視覚的な補助を与えることが出来ると考えられる。つまり，教

示用の動画と，学習者が発声した際の口唇の動きを記録した動画との間で発話速度の同期をとり，同時に学習者に提示することで学習者に正しい口唇の動作を認知させることを考えている。

2.3.1 音声合成部

GalateaTalk は形態素解析システム 茶筌 [6] を用い漢字仮名混じり文で表記された日本語テキストを形態素解析し，HMM をベースにした音声合成を行なう。

対話システムを構成するため，日本語学習を行う際に疑似教師のようなインタフェースを作成する。本研究ではライセンスフリーのソフトウェアツールキットとして配布されている擬人化音声対話エージェント Galatea を用いる。Galatea は音声認識，音声合成，顔画像合成の 3 つの基本機能を統合し，対話制御のもとでユーザと対話するエージェント，及びその開発環境を提供するものである。GalateaTalk による合成音声は，学習者と対話する際に補助的に用いることを考えている。教示用音声には録音した音声を用いる。

Galatea を使用することで，より教師らしいヒューマンインタフェースの作成する。つまり，総合統括部で学習状況や発話により決定されたアドバイスを機械的に与えるのでなく，人間的に提供することで学習者が学習しやすい環境を提供していく。

2.3.2 画像認識部

本研究ではクラスタシステムを用いた高い処理能力を活かして，使用者の口の動きを画像認識技術を用いて解析，学習者に結果を提示することを想定している。画像認識部では発話時の画像を収録し，教師の発話音声との DTW(Dynamic Time Warping) を行ない，画像及び音声の同期をとって再生することで発話訓練に利用することを考えている。

3. まとめ

コンピュータの語学教育への活用は，人間的判断を伴う語学教師の仕事を完全に代行することは出来ない。現段階ではあくまでも学習の一部を補う補助的道具として，また学習方法や教材の種類や効率を高めるために利用されている。現在，外国語学習システム [7][8] や，日本語の書取学習を補助するシステム [9] など，さまざまな研究機関で語学学習システムが構築されている。本研究が提案するシステムでは，日本語特有の発話現象である長母音の知覚訓練 (例えば「おばあさん」と「おばさん」ように，日本語で長母音と単母音を区別しないと問題が多い) や，母音の無声化 (日本語の狭母音（/i/と/u/）は前後を無声子音に囲まれると無声化する現象) に対する知覚訓練について，充分考慮したものとする計画である。

4. 今後の予定

現在，実際にクラスタシステム上で動作し，学習者にテキストや画像を掲示して学習者の応答を認識し結果を提示することが出来る。来春までの課題として，

- 約 1ヶ月分の初級日本語学習者テキストを教材として作成する
- 教師用音声として利用する音声の録音
- 発話画像との同時提示や口唇部分の拡大画像の提示
- 完全な初習者に対して情報を提示する場合を考え，画像・アニメーション・受講者の母国語での説明を行なう。当面はUTFを利用し韓国語を作成することを考えている

を挙げている。

参考文献

[1] http://www.moj.go.jp/
[2] http://www.bunka.go.jp/index.html
[3] http://julius.sourceforge.jp/
[4] http://hil.t.u-tokyo.au.jp/ galatea/
[5] http://www.nime.ac.jp/tokutei120/ 05publication/01/a02/a02_09.pdf
[6] http://chasen.aist-nara.ac.jp/
[7] 峯松信明，仁科喜久子，中川聖一：”外国語学習用読み上げ音声データベース”日本音響学会誌，**59**，pp.345-350，(2003)
[8] 中川聖一，牧野正三，壇辻正剛 :発声言語処理技術を用いた語学学習システム日本音響学会誌，**59**，pp.337-343，(2003)
[9] http://sp.cis.iwate-u.ac.jp/sp/lesson/j/