社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2002-113, WIT2002-53(2002-10)

聴覚障碍者のための口形つきアニメーションの教材に関する検討

喜田村 朋子† 荒井 隆行† Pamela Connors‡

† 上智大学理工学部電気・電子工学科 〒102-8554 東京都千代田区紀尾井町7番1号
‡ Tucker-Maxon Oral School 2860 SE Holgate Blvd. Portland, Oregon 97202, U.S.A.
E-mail: † toko@splab.ee.sophia.ac.jp

あらまし CSLU (Center for Spoken Language Understanding at the Oregon Graduate Institute) において開発されてきたCSLU Toolkitの Baldi と呼ばれる口形つきアニメーションによる教材が、口話によるろう学校 Tucker-Maxon (Portland, Oregon)において使われている。このような教材は、日本においてあまり使われていない。そこで日本語のエージェントを実現するために、その教材に含まれる Mexican Spanish の viseme を日本語の viseme に対応させることを試みた。さらに日本において使用するときに考えられる可能性について、特に以下の点について検討した。

(1) 口形つきアニメーションによって、聴覚情報だけでなく視覚情報も得られること。
(2) 口形つきアニメーション・音声合成・音声認識・画像の組み合わせにより、比較的自由に教材を作成できること。

検討の結果、国内で発音訓練や読話訓練、語彙獲得のために有効な教材になることが示唆された。

キーワード アニメーション、聴覚障碍、読唇、発音訓練

A study on education tools with an animated talking agent for the hearing impaired

Tomoko KITAMURA† Takayuki ARAI† and Pamela Connors‡

† Department of Electrical and Electronics Eng., Sophia University 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan
‡ Tucker-Maxon Oral School 2860 SE Holgate Blvd. Portland, Oregon 97202, U.S.A.
E-mail: † toko@splab.ee.sophia.ac.jp

Abstract An education tool with an animated talking agent named Baldi, that has been developed at CSLU (Center for Spoken Language Understanding at the Oregon Graduate Institute) is used at Tucker-Maxon (Portland, Oregon), a school for deaf children using oral language. Tutors like the CSLU Toolkit have not been used in Japan. In order to create a Japanese language animated agent, we have tried to map visemes of Mexican Spanish phonemes to the corresponding visemes of the Japanese phonemes. In this report, we proposed the ways in which these resources can be used effectively in Japan. We focused on the following areas:

(1) We can obtain not only audible information but also visible information from an animated talking agent;
(2) By combining an animated talking agent, speech synthesis, speech recognition, and images, we can make any kind of tutor with the CSLU Toolkit.

The result of the examination showed that the CSLU Toolkit could be effective for training pronunciation, speech reading, and the acquisition of vocabulary in Japan if there were a Japanese voice, text to speech system, speech recognition and facial animation in Japanese within the CSLU Toolkit.

Keyword animation, hearing impaired, speech reading, training of pronunciation

1. はじめに

聴覚障碍者の聴覚および口話によるコミュニケーション能力を高めるための教材は、これまでにもいくつか市販されてきた。聴覚障碍者にとっては、他人の発する音声と自分の発する音声を耳で聞いて比較・判断し、フィードバックすることが難しい。そこで正しい発音方法を体得するための教材としてエレクトロパラトグラフィー(EPG)や舌センサなどを用いた練習機が使われていた。しかし、これらは専門的な取り扱いが必要な上に、大変高価なものである。それに対し、アニメーションを用いてゲーム感覚で楽しく発声・発音の練習ができ、容易に操作ができる安価なソフトウェアが作られてきた [1-2]。「あいちゃんの手」を例に

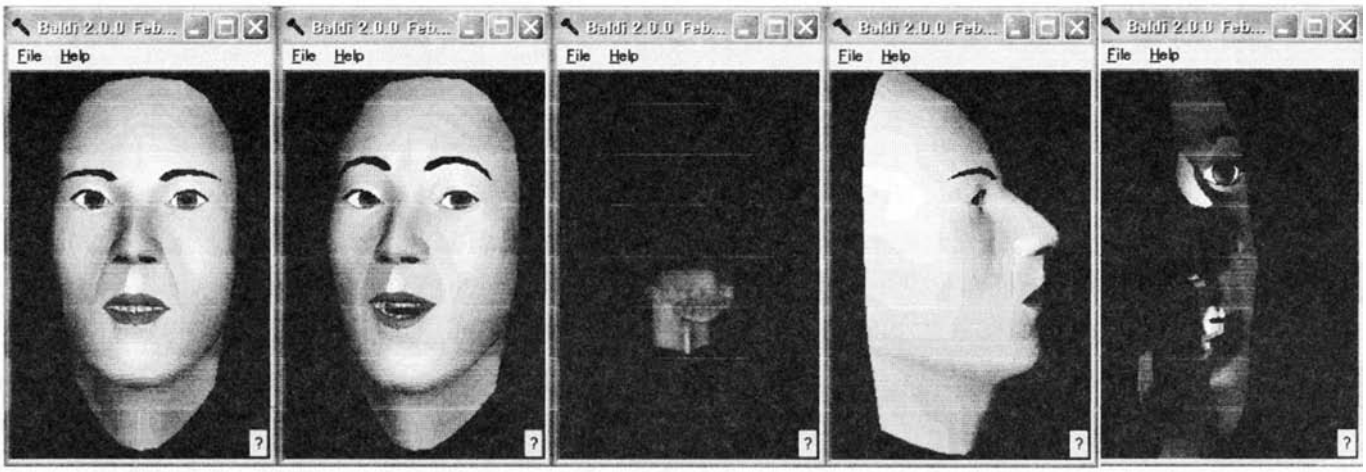

図 1: 左より顔（標準）・顔（Happy）・顔の表面を透過させたもの・横顔にしたもの・顔の断面図

挙げると、マイクロユニット、インジケータ、アクティブ・ハンド、ソフトウェア、操作パッドの５点という簡単なシステム構成で利用することができる [1]。訓練メニューとしては気息、発声、母音発音、韻律などが含まれ、これらを視覚と手の触覚を利用し聴覚も活用しながら、発声・発音の学習ができるように工夫がなされている [1]。

他方で、現在コンピュータと人間とのより円滑なコミュニケーション実現のために、リアルに話し言葉を発するエージェントを作り出すインターフェイス技術の研究が盛んに行われている。これは、あたかも人間同士が対話をするように、コンピュータと人間とのコミュニケーションを実現するグラフィカルユーザインターフェイスである。電子レンジについて消費者に情報を与えるために開発されてきた Olga プロジェクト [3] や、ストックホルム諸島における船の交通情報に関する情報を与える WAXHOLM [4]、ストックホルムにある施設等の情報を与える August プロジェクト [5] という対話システムなどに代表される。これらは、コンピュータに関する知識の少ない利用者が、エージェントと話し言葉による対話をしながらコンピュータを利用しやすくすることを目的としている。また、エージェントを用いることによる発話中の視覚情報（唇・舌・顎の動き）の重要さにも着目している。特に雑音環境のように聴覚情報の減少している中において、視覚情報が加わることによって話の理解の明瞭性が上がる。口形つきのエージェントによる視覚的情報を加えると、人間の自然音声および合成音声の明瞭度が増加することが分かっている [6]。また音声によるコミュニケーションは、発話者の表情、感情、ジェスチャーによって豊かなものとなる [7]。さらに、こういった視覚情報は、聴覚情報の欠損しがちな聴覚障碍者にとっても重要なものである。従って聴覚的にも視覚的にも正しく発話することのできるエージェントは、高齢化社会において増加が予想される聴覚障碍者にとってとても可能性のあるものである。なぜならば聴覚障碍者にとって、視覚的に発話を読みとる「読話」や、残存聴力を活用して言葉を聴きとる訓練のできるエージェントは非常に意味のあるものだからである。ここで「読話」とは、音声を産出する時の発声発語器官の動きを外から視覚的に捉えて音声を認識しようとするものである [8]。

そこで、口形つきのアニメーションを用いたインターフェイス技術が、聴覚障碍者の語彙獲得・読話訓練・発音訓練のためのレッスン教材として口話によるろう学校 Tucker-Maxon (Oregon, Portland) で使用されている例があるので、それについて紹介する。

2. CSLU Toolkit

1993 年より、CSLU（Center for Spoken Language Understanding at the Oregon Graduate Institute）において使いやすく、話し言葉に基づく様々な機能を含んだソフトウェアが開発されてきた。それが CSLU Toolkit である。この Toolkit はテキスト音声合成、音声認識、アニメーションの各技術が組み合わされたものである。そして基礎研究のサポートや、ユーザインターフェイスと話し言葉システムに関した教育・開発のために作られた [9]。

この CSLU Toolkit に含まれる機能の中でも、聴覚障碍者の言葉のレッスン教材のために特に使われる機能について以下に述べる。

2.1. 動作環境

現状では、CSLU Toolkit を導入するにあたり、必要とされる動作環境は次の通りである。

(1) ハードウェアの必要最低の条件

- Pentium Pro 200 MHz 以上であること
- 128 MB 以上の RAM を有すること

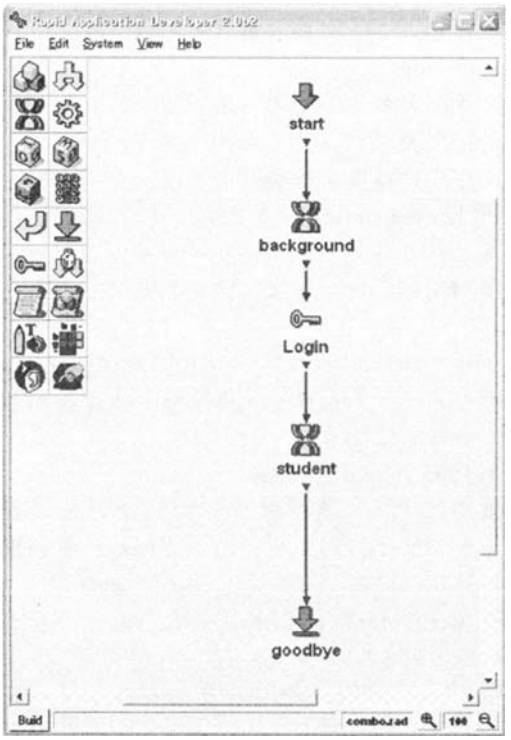

図 2: RAD の例

- 8MB RAM のビデオカード, hardware Open GL accelerator
- Windows に含まれるオーディオカード

(2) ソフトウェアの動作環境

- Windows 95, 98, ME, NT, 2000

(3) CSLU Toolkit のダウンロード

http://cslu.cse.ogi.edu/toolkit/download/index.html

2.2. Facial animation

PSL(Perceptual Science Laboratory at the University of California, Santa Cruz)で開発された Baldi と呼ばれる 3D アニメーションが用いられている。CSLU Toolkit で用いられているこのアニメーションの特性は以下の通りである[10]。

- 唇・舌・顎・顔の動きと、合成音声および録音された人間の自然発話音声が自動的に同期する。
- 顔の表面（肌）を透過させ、歯や発話中の舌の動きを見ることができる。
- 顔の角度を変えることによって、発話の様子を様々な角度から見ることができる。
- 顔の表情を変えることができる。
- 顔の断面図を見ることができる。

顔の表情を変えたもの、肌を透過させたもの、顔の角度を変えたもの、顔の断面図を見られるようにしたものについては、図 1 に示す。

2.3. Speech synthesis

CSTR (The Centre for Speech Technology Research University of Edinburgh)で開発された Festival と呼ばれるテキスト音声合成システム[11]が用いられている。CSLU Toolkit で用いられているこのシステムの特性は以下の通りである[10]。

- 男声・女声、American English・Mexican Spanish を含む数種類の合成音声を作ることができる。
- テキストから、文章の文節ごとに適切な持続時間およびピッチ周波数や大きさなどの韻律的な要素をうちだし、合成音声に変換する。その際、diphone または unit-selection concatenative synthesis が使われる。
- 合成音声の発話スピードやピッチ周波数を変えることができる。

2.4. Speech recognition

人工的ニューラルネット(ANN)型識別器、隠れマルコフモデル(HMM)および音節モデルが使われている。大人の声または子供の声を基準に設定し、教材の中で認識させることが可能である[10]。

2.5. Rapid Application Developer（RAD）

Toolkit の中には、教材を作るダイアログが含まれている。これらは、RAD と呼ばれるものである。RAD を立ち上げると、空きのキャンバスとツール表が表れる。このときのウィンドウを図 2 に示す。この表のツール一つ一つはオブジェクトを表し、toolkit に含まれる技術、例えば音声合成・音声認識・アニメーション(Baldi)などが使えるようになっている。また字幕や画像の表示・非表示も自由にコントロールすることができる。教材を作成する際、ツール表からキャンバスにオブジェクトを drag-and-drop し、並べ、さらにそれらを線で結ぶ。そして Baldi のセリフなどを指定されたところに入力したりする[12]。教材作成者の案によっては、オブジェクトの並べ方や工夫次第で、様々なスタイルの教材を作り出すことができる。その例については後に示す。生徒がこのレッスンを実行すると、作成された教材に沿って、Baldi の話を聞き、あるいは反応することができる。

3. Tucker-Maxon Oral School（TMOS）

Tucker-Maxon は、Oregon 州 Portland にあり、50 年以上の歴史を持つ。人工内耳あるいは補聴器を装用した、乳児から高等学校にいたる高度聴覚障碍児のための私立小学校である。ここでは手話を使わず、口話によって教育が行われている。ここで、Baldi は、生徒たちにとって補足的な会話のパートナーとなるだろうと期待され、導入された [13]。1997 年 9 月より、CSLU と PSL は NSF Challenge grant を受け、3 年間のプロジェクトで TMOS との共同研究を通し、Baldi と CSLU toolkit を聴覚障碍児のための言葉の学習アプリケーシ

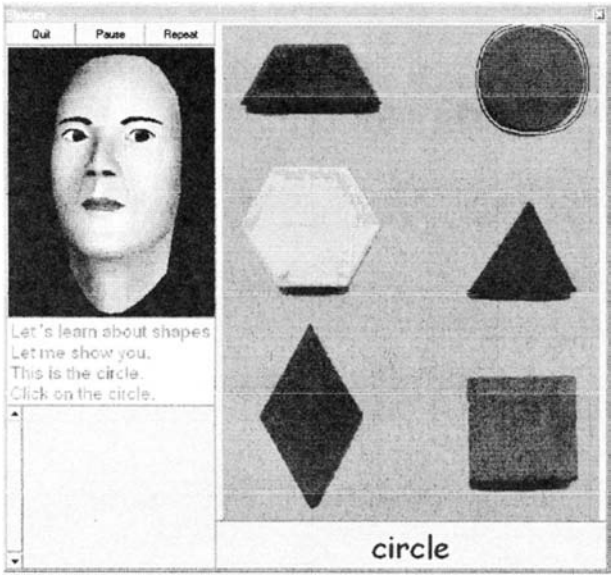

図 3：Vocabulary tutors の例

ョンとして開発してきた[13]。これを通して RAD は、より使いやすくわかりやすいものに改善されてきた[12]。そして TMOS で教える先生たちの意見が取り入れられ、後に述べる Vocabulary tutor が作られた。この章では Vocabulary tutor について説明した後、実際にどのような教材が作られ使われていたかを述べる。

3.1. Vocabulary tutors (VT)

Vocabulary tutors (VT) と呼ばれるウィザードを用いることによって、数十分の短時間で簡単に言葉のレッスン教材を作ることができる。この VT は、Baldi の顔と、合成音声または録音された人間の自然発話音声と、画像によって構成される。VT を立ち上げた際のウィンドウについて図 3 に示す。教材作成者は、生徒に覚えてもらいたいアイテムを画像から選び出して線で囲み、そのアイテムの名前を入力していく。この VT は 8 ステージによって構成されている。ただし作成者の意向によって任意のステージを省いたり加えたりすることができる。各ステージの大まかな流れを下に示す。

(1) Pretest

言葉のレッスンが始まる前の生徒の保有語彙数を調べるためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Click on the <画像に含まれるアイテムの名前>"

生徒: マウスを使って、そのアイテムを画像の中から探し、クリックする。

VT の作成者によって決められたそれぞれのアイテムは線で囲まれ、マウスが各アイテムの上を通るとその部分がハイライトされるようになっている。この動作が、アイテムごとに繰り返される。

(2) Presentation

各アイテムの画像とその名前（単語）を視覚的および聴覚的に関連づけさせるためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Let me show you."

画像にある各アイテムがハイライトされる。

Baldi: "This is the <画像に含まれるアイテムの名前>."

そのアイテムの名前のスペルも同時に表示される。そしてそのアイテムをクリックするように指示され、生徒はその指示に従う。

(3) Perception

生徒たちにとって各アイテムの画像とその名前が結びつくまで反復練習する。ここで行われる動作の例は以下の通りである。

Baldi: "Let's practice. Click on the <画像に含まれるアイテムの名前>."

生徒: 正しいと思われるアイテムをクリックする。

Baldi: "Good/ Sorry."

生徒の反応が正しいか間違っているかによってどちらかの返事がなされる。またニコニコマークあるいは残念マークが表示される。

全てのアイテムにおいて生徒が正しく反応できるようになれば、次のステージに進む。

(4) Word ID

各アイテムとその名前の表示を関連づけるためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Click on the word below the picture when I show it to you."

生徒: Baldi に示されたアイテムの名前を下に並べられた単語の列から選び、クリックする。

これに対する Baldi の反応は(3)の場合と同じである。

(5) Spelling

各アイテムの名前を正しいスペルでタイプできるようにするためのステージ。ここで行われる動作の例は以下のとおりである。

Baldi: "Type the word that I show you."

生徒: Baldi に示されたアイテムの名前のスペルをタイプする。

これに対する Baldi の反応は(3)の場合と同じである。

(6) Imitation

各アイテムの名前を、Baldi に続いて正しく発することができるようにするためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Repeat the word after me. This is <画像に含まれるアイテムの名前>. Say <画像に含まれるアイテム

の名前>."

生徒:示されたアイテムの名前を Baldi に続いて発する。

生徒の発した音声は録音され、再生される。

(7) Elicitation

各アイテムの名前を正しく発することができるようにするためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Let's practice saying the words. What is this?"

生徒: 示されたアイテムの名前を発する。

生徒の発した音声は録音され、再生される。続いて Baldi の声による正しい発音も再生され、生徒は双方を聴き比べることができる。

(8) Post test

言葉のレッスンが終わった後の、生徒の獲得語彙数を調べるためのステージ。ここで行われる動作の例は以下の通りである。

Baldi: "Click on the <画像に含まれるアイテムの名前>."

生徒: マウスを使って、そのアイテムを画像の中から探し、クリックする。

この反応に対する Baldi の返事は(3)の場合と同じである。ただし、各アイテムについてトライできる回数は 1 回ずつである。

全ステージにおいて、字幕を表示するかしないかを決めることもできる。また、Baldi の顔の形態を変える(顔の表面を透過させるかどうかなど)こともできる。そして各ステージにおける結果をデータログすることができる。これは HTML 形式でまとめられ、生徒の得た正解数や録音された音声を、指導者があとで確認することができる。

3.2. TMOS における教材例

TMOS で使われていた教材の例をここに挙げる。これらは、すべて RAD および VT によって作成可能である。ただし、いずれも実際に学校の授業や家庭において、先生や家族などの指導者による指導後、補足的に復習することを主な目的として使われていた。

例1) 語彙を増やすための絵カード代わりに使う。例えば、アメリカの州名を覚えるために使うなど。

例2) 絵本の代わりに使う。絵本の絵をスキャナで読み込み、それを表示しながら、先生の声と Baldi の口の動きを同期させて読み上げる。

例3) 子音の聞き取り訓練に使う。例えば、"bag" と"wag"のような minimal pair の単語を聞き取るなど。

例4) 正しい発音によって場面が進むようにする。例えば、ピザの注文の場面を想定し、「ピザの大きさは大きいのと中ぐらいのと小さいのとどれが良いですか?」などという質問に生徒が口頭で答え、正しく認識されれば次の場面に移るようにするなど。(ただし、これは RAD によって作成)

4. CSLU Toolkit の日本語への応用例

最初に、既存の CSLU Toolkit のプログラムを用い、Baldi に日本語のルールに従って口を動かすことを実現した。現在手に入る CSLU Toolkit は American English と Mexican Spanish の音声合成データのみ含んでいる。American English に比べ Mexican Spanish と日本語の構音が似ていることに着目し、Mexican Spanish と日本語の子音および母音の viseme を対応させた。Mexican Spanish に含まれる子音で、日本語でも似た構音をするもの、例えば歯茎破裂音/t/などは、そのまま使用した。日本語にはあるが Mexican Spanish にない音、例えば日本語の声門摩擦音/h/は、Mexican Spanish の軟口蓋摩擦音/x/のように比較的近い構音をするものを代用している。こうして、日本語による自然発話音声を録音し、その音声データに Baldi の口の動きを同期させるためのテキストから viseme へのプログラムを作り替えた。これは、RAD による教材で使うことができる。

これを用いた場合、どういった使い方が効果的であるか、日本の言語聴覚士やろう学校の先生に案を出していただいたので、それをまとめる。

まず、CSLU Toolkit を、語彙獲得よりも読話力向上のために用いることにポイントをおくとするならば、中途失聴者や高齢難聴者を対象にすることが適切と思われる。その理由として次のことが挙げられる。コミュニケーションにおける聴覚併用読話率と読書力との間には相関関係があることがわかっている [8,14]。つまり、言語理解力しいては言語能力が聴覚障碍者の読話能力を規定する一つの重要な因子となっている[14]。なぜならば、読話は発声発語器官の動きの知覚をもとに、音韻ないし音節を推測し、さらに語や統語などの言語能力、話し手の非言語的伝達の受信、場面や状況的手がかり、さらには残存聴力による聴覚的情報、相手についての知識など、きわめて多くの情報を総合して、"話" を推測するからである [8]。このことより、CSLU Toolkit を、中途失聴者や高齢難聴者のようにすでに言語能力の習熟している人たちのための読話力向上の訓練に活用できることが言える。

ここに、中途失聴者が読話の技術を習得するための過程と、そこで CSLU Toolkit がどのように活用できるかを示す。

(1) 単音節レベルで読話の練習をする。

(2) 単語レベルで読話の練習をする。

(1)(2)より、日本語の語音と口形の関係の規則を体得していく。

(3) 話し手の表情、しぐさや話しているときの状況などの非音声的要素から話の内容を予測することに慣れる。そして、様々な人の色々な話し方に慣れていく。

中途失聴者には、これまで健聴者であり、自分の構音を意識せずに発話してきた者が多い。そのため、外部から見ることのできる発声発語器官の動きは限られており、見える動きが同じようでも、見えない部分で異なって発音される音は区別ができない[8]ということを理解することが難しい。つまり、"同口形異音"が多数存在することが理解できないのである。これは読話講習会などの限られた時間内に習得することには限度がある。従って家に帰った後、鏡を使って自分の口形を見ながら独習することも必要とされる。しかし、この方法では、自分で発しようとしている単語が既に自分でわかっているため、読話の練習にならないという問題が起きる。そこで、CSLU Toolkitを用いてBaldiにいろいろな単音節または単語を言わせ、それを読みとる練習をするといった、読唇への活用方法が提案できる。

また、TMOSで実施されているように使うこともできる。つまり、CSLU Toolkitを言語能力の安定していない年齢の聴覚障碍児のための、指導者や家族と共に学習したことを復習する補足的な学習としての語彙獲得のために活用するのである。またBaldiの表面（肌）を透過させ、舌の動きと口蓋との対比を観察し、構音の練習をするためのモデルとして用いることもできる。そして相手がコンピュータであることを利用し、何度もBaldiに繰り返ししゃべらせることもできる。

5. まとめ

本論文において、口形つきアニメーションが含まれたCSLU Toolkitについて紹介した。そして口形のついているアニメーションによって、聴覚情報だけでなく視覚情報も得られることを説明した。このCSLU Toolkitを用い、アニメーションおよび合成音声または録音された人間の自然発話音声と音声認識と画像の組み合わせによって、作成できる教材には様々な可能性がある。このCSLU Toolkitを、我々は日本語のルールに従い、また録音された人間の自然発声音声に同期させてBaldiが口を動かすことのできるようにした。これを用いることによって、日本においても聴覚障碍者に対して有効な教材が作れることを提案したい。

6. 謝辞

CSLU Toolkitについてご指導くださったTucker-Maxonの皆様に厚く御礼を申し上げます。また、日本における言語教育についてご指導くださった帝京大学名誉教授田中美郷先生、ノーサイドの芦野聡子先生をはじめとする皆様に感謝申し上げます。最後に、CSLU Toolkitの使用例について共に検討くださった神奈川県聴覚障碍者福祉センターの岩崎友子先生、山内浩一先生、そして横浜市立聾学校の長谷房代先生にも厚く御礼申し上げます。

文　献

[1] 中村敬和, 昆昭彦, 青木功, 浅輪晃一, "聴覚障碍児用 発声練習システム「あいちゃんの手」," pp.98-105, 音声言語情報処理, 1998.

[2] Speech ViewerIII, 日本IBMアクセシビリティーセンター
http://www-6.ibm.com/jp/accessibility/soft/hearing.html#navskip

[3] J. Beskow, K. Elenius, and S. McGlashan, "Olga - A dialogue system with an animated talking agent," EUROSPEECH, pp. 69-72, 1997.

[4] R. Carlson and B. Granström, "The Waxholm spoken dialogue system," Palková Z, ed. Phonetica Pragensia IX. Charisteria viro doctissimo Premysl Janota oblata. Acta Universitatis Carolinae Philological, pp. 39-52, 1996.

[5] M. Lundeberg and J. Beskow, "Developing a 3D-agent for the August dialogue system," AVSP, 1999.

[6] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K. Spens and T. Ohman, "The teleface project multi-modal speech-communication for the hearing impaired," EUROSPEECH, 1997.

[7] D. W. Massaro, "Perceiving Talking Faces," Speech Perception to a Behavioral Principle, MIT Press, 1998.

[8] 坂本幸, 恩納亮子, "読話における口形の弁別力と個人差の影響について," 宮城教育大学紀要第33巻, pp.157-168, 1998.

[9] R. A. Cole, "Tools for research and education in speech science," ICPhS, pp.1277-1280, 1999.

[10] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, Ed. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro, and M. Cohen, "Universal speech tools: The CSLU Toolkit," ICSLP, pp. 3221-3224, 1998.

[11] A. W. Black, P. Taylor, and R. Caley, "The Festival Speech synthesis System," System documentation, Edition 1.4, for Festival Version 1.4.0, 1999. http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html

[12] P. Connors, A. Davis, G. Fortier, K. Gilley, B.Rundle, C. Soland, and A. Tarachow, "Participatory design: classroom applications and experiences," ICPhS, pp.1285-1288, 1999.

[13] P. Stone, "Revolutionizing language instruction in oral deaf education," ICPhS, pp.1281-1284, 1999.

[14] 田中美郷, 進藤美津子, 本宮敏司, "テレビを用いた読話テストと高度聴覚障碍者のコミュニケーション能力について," Audiology Japan Vol. 16, pp.109-119, 1973.