Vol.2015-CE-131 No.2
2015/10/10

専門科目英語講義の字幕付き動画作成を支援するシステムの設計

森山 聡[†1]　宇佐美　格[†2]　小川　亮二[†2]　八木 佑侑季[†2]
戸辺　義人[†1]　鷲見　和彦[†1]

近年，大学における教育においても英語による講義が広がりつつある．我々は，受講生に対して内容の理解を助けるために，英語で話された講義動画から，Sphinx または docomo 音声認識エンジン・docomo 翻訳エンジンを用いた音声認識・英日翻訳を経て，日本語・英語の字幕付き講義動画を生成するシステム SACMI を開発した．多くの場合，科目担当者は内容の専門家ではあっても英語を母語としない講師であるため，一般的な音声認識システムで話される内容を認識するのは困難となる．また，講義では専門用語が多く出現するため，一般的な翻訳システムでの翻訳も困難である．そこで SACMI では，誤った認識結果，翻訳結果を容易に編集して，作業時間を簡略することを狙う．本稿では，SACMI の概要，設計，実装と評価結果について述べる．

Design of a System to Assist Generation of Videos with Subtitles for Lectures in English

SATOSHI MORIYAMA[†1]　ITARU USAMI[†2]　RYOJI OGAWA[†2]　YUKI YAGI[†2]
YOSHITO TOBE[†1]　KAZUHIKO SUMI[†1]

Recently, lectures in English are being provided at universities. We developed a system called SACMI, which generates videos with subtitles in Japanese and English from lectures in English through speech recognition and translation into Japanese with Sphinx or docomo speech recognition API and docomo translation API in order to help students understand lectures. In many cases, conventional speech recognitions misrecognize voice data because professor is not usually a native English speaker. Moreover, conventional translation systems also mistranslate because there are many technical words in lectures. That is why, SACMI simplifies revising the results of speech recognition and translation, and shorten of work hours. In this paper, we describe overview of design of SACMI, the implementation, and the evaluation.

1. はじめに

近年，企業活動のグローバル化に対応できる人材育成や，大学の国際化と留学生受け入れ推進に伴い，大学における理工系専門科目講義を英語化していこうとする動きがある．しかし英語を母語としない受講者にとって，講義内容のすべてを英語で受講することは簡単なことではなく，理解度の低下が懸念される．英語の話すスピードを遅くしたり，講義内容を簡略化したりすることで講義自体の理解度は改善できるかもしれないが，理工系の専門講義の趣旨を逸脱してしまう．

そこで，講義自体の質を落とさずに講義を英語化する手段として，講義中に撮影した動画に英語・日本語字幕を付けた予習・復習教材を作成する仕組みが必要である．従来，このような字幕を作成するためには，話者の英語の口述を筆記する音声認識，英日翻訳，話者の英語と同期した字幕の生成と表示のための映像編集の３つの作業をすべて人の手により行う必要があった．この作業には，講義時間の何十倍もの時間を必要としており，講義準備の負担が大きすぎた．そこで我々は話者の音声認識，英日翻訳，字幕の生成，字幕付き動画の再生を行うことができるシステムの研究 SACMI (Study for Annotated Course Material for Internationalization) を行った．SACMI ではカーネギーメロン大学が開発する音声認識エンジン Sphinx[1)]を用いることにより，従来の音声認識エンジンでは認識することが困難であった，英語を母語としない話者の英語の認識精度を向上することを狙っており．また，NTT docomo 社が提供するクラウド型翻訳エンジンを用いることにより，ユーザ辞書登録機能を利用して一般的な日常会話向けの翻訳エンジンを理工系の専門用語にも対応できるようカスタマイズして翻訳精度を向上することを狙っている．さらに，従来すべて人の手で行っていた音声認識・英日翻訳・映像編集のプロセスを簡略化させることにより，作業時間を短縮させ，教材作成効率を向上させることを狙っている．

以下 2 章では関連研究について述べる．3 章では設計と実装について述べる．4 章では評価実験について述べ

†1 青山学院大学理工学部情報テクノロジー学科
College of Science and Technology Aoyama Gakuin University
†2 青山学院大学大学院理工学研究科
Graduate School of Science and Technology Aoyama Gakuin University

る．5 章では評価実験の考察について述べる．6 章では本研究の結論を述べる．

2. 関連研究

講義動画から英語を認識させる際，英語を母国語としない話者の英語認識率は作業全体の効率に大きく影響する．朝川らは，英語の母音の構造的表象に着目し，日本語と英語に共通する母音のみで英語を発音した状態から，より英語らしい母音を発音するよう矯正していく過程を可視化している．また，英語の強勢弱勢母音についても着目し，音素間の距離行列からクラスタリングおよび音声構造の生成を行い，母音の強弱がはっきりするほど大きくなる韻律的音声構造のサイズが英語の習熟度と比例することを示している [2)]．

大橋らは，一般的な前後音素に着目してクラスタリングを行い，日本人英語と日本人日本語をマルチパス置換する認識モデルを構築している．これに加え，発話される単語のスペル母音を状態分割決定木で表現したものを利用することで，スペル母音依存の高い日本人英語の認識率向上を実現している．同時に，習熟度によって最適な認識モデルを利用できる汎用性も実現している [3)]．

SACMI は，講義動画の翻訳という目的で利用されるため，汎用的な認識モデルよりもむしろ講義教員一人ひとりに合った特定話者認識モデルを利用することが望ましい．SACMI において音声認識ツールとして利用している Sphinx は文法チェックを行わないが，認識対象となる人物に合わせた認識モデルをカスタマイズ可能な仕様となっているため，発話者ごとに最適な認識モデルを構築することが可能になっている．

3. 設計と実装

SACMI はあらかじめ撮影しておいた英語での講義ビデオに対し，字幕を表示させるための字幕ファイル生成支援システムの研究プラットフォームである．SACMI は，音声認識，英日翻訳，字幕の生成・表示の 3 つの段階で構成される．従来は，翻訳作業者が実際に音声を数回繰り返し聞いて文章に起こし，日英翻訳や字幕の表示時間などの編集作業をしていたため，多くの作業時間を費やしていた．SACMI は，これらのプロセスは簡略化することを可能とした．SACMI における音声認識から字幕ファイル生成までのプロセスは図 1 のとおりである．本章では，SACMI の設計・実装について述べる．

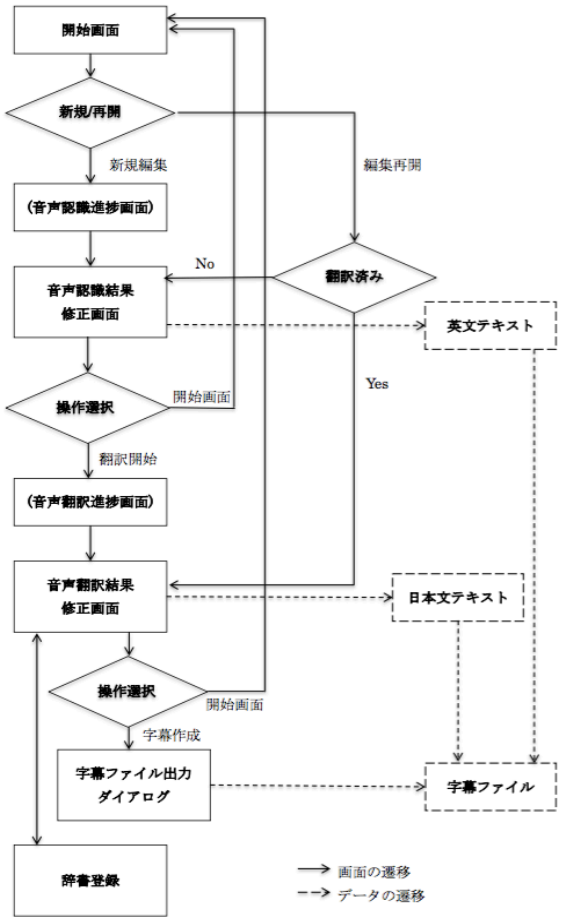

図 1　字幕ファイル生成までの状態遷移図
Figure 1　State transition diagram to create subtitles files.

3.1 事前処理

SACMI を使用するための事前処理として，講義ビデオから音声認識が可能となる単位音声ファイルにセグメント化する．その際の音声ファイルは wav ファイルである必要がある．wav ファイルのフォーマットはサンプルレートが 16kHz，ビット/サンプルが 16 ビット，オーディオチャンネルはモノラルとする．次に，音声を発話(文)毎に切り出し，1.wav, 2.wav, 3.wav, …, n.wav （1≦発話数≦n）のように wav ファイルの名前を 1 から分割ファイル数分，順番に名前を付ける．音声の切り出し処理を行う理由は，音声認識の精度向上と，字幕ファイル生成時にそれぞれの文の発話・終話時刻を音声ファイルの累積時間を計算して生成するためである．

3.2 音声認識

音声認識では，事前処理にてセグメント化した各音声ファイルの音声を英文の文字として取得する．音声認識エンジンは，編集者がクライアントアプリケーションにて，docomo 音声認識エンジンおよび Sphinx から選択す

ることを可能とする．docomo 音声認識エンジン選択時には，NTT docomo が所有する音声認識サーバとインターネットを介して通信し，音声認識結果を受信する．Sphinx 選択時には，編集者が事前処理の過程で生成したローカル内に保持する音声ファイル群を，我々が構築した Sphinx サーバに送信し，その音声認識結果を受信する．音声認識において Sphinx を用いる場合，講義ビデオ内に登場する専門用語の学習や，英語を母語としない話者への適用を行うことを可能とした．クライアントは音声認識結果を受信すると，音声認識結果ファイルを生成し，クライアントアプリケーション内で保持する．音声認識結果ファイルの中身は図 2 のようになっており，1 行目が発話開始時刻，2 行目が終話時刻，3 行目が音声認識で得られた英文を表している．

音声認識終了後，クライアントアプリケーションは，図 3 の音声認識結果修正画面に遷移する．編集者が生の音声を聞きながら編集作業ができるように，画面左側に配置した音楽プレーヤにより修正中の音声ファイルを再生することを可能とした．画面右側には，音声認識エンジンで得られた英文を表示し，誤認識があった場合には手作業で修正することが可能である．ここで修正された情報は，音声認識結果ファイルに自動で上書きさせる．

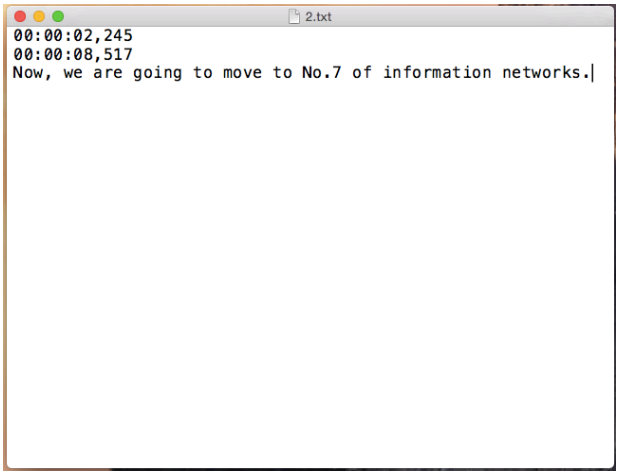

図 2 音声認識結果ファイル

Figure 2 The file of speech recognition result.

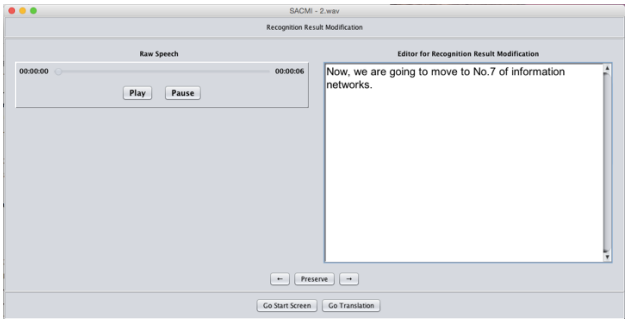

図 3 音声認識結果修正画面

Figure 3 The screen of speech recognition result.

3.3 英日翻訳

英日翻訳では，前述の音声認識後，編集者によって修正された英文テキスト群から，日本語への翻訳結果を取得する．まず，音声認識・修正後の英文テキスト群を NTT docomo が所有する英日翻訳サーバへとインターネットを介して送信し，その結果として日本語に翻訳された結果を英日翻訳結果テキストとして受信する．受信した英日翻訳結果テキストを翻訳結果修正画面にて，編集者が順次修正していくことができる．ここで修正された情報は，対応する英日翻訳結果テキストに自動で上書きする．英日翻訳結果ファイルの中身は図 4 のようになっており，1 行目が発話開始時刻，2 行目が終話時刻，3 行目が音声認識・修正後の英文，4 行目が翻訳結果の日本語文を表している．

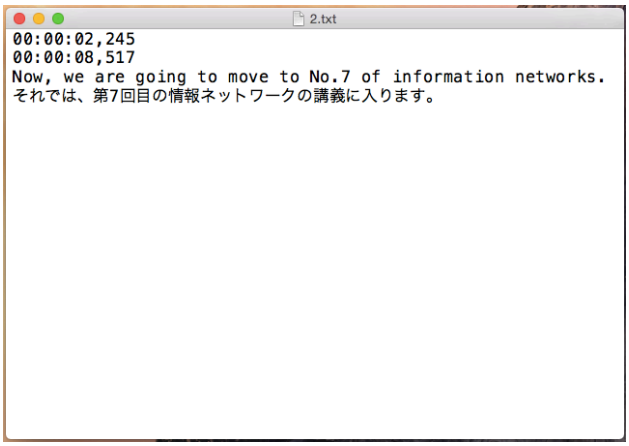

図 4 英日翻訳結果ファイル

Figure 4 The of translation result.

また，本システムでは，翻訳エンジンにて翻訳誤りがあった英単語に対して，本来意図していた翻訳結果を記録するため，図 5 の画面を用いて，docomo 英日翻訳サーバが有するユーザ辞書登録機能を活用する．SACMI では，この辞書登録機能を使用するための辞書登録インタフェースを用意し，辞書に登録された単語からいくつかの候補を提示することを可能とする．このインタフェースを図 6 に示す．

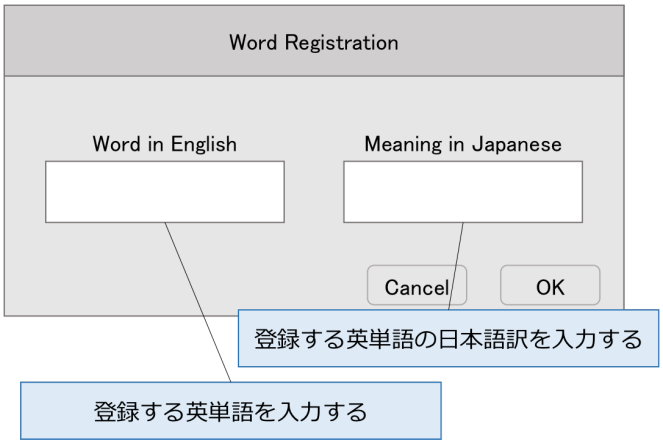

図 5 新しい単語の登録画面

Figure 5 The registration screen of a new vocabulary.

図 6 辞書登録された単語訳の候補画面

Figure 6 The screen for showing candidates.

3.4 字幕生成

字幕生成では，上記の工程で生成した音声認識・修正後の英文，英日翻訳・修正後の日本語文，およびそれらが発話された時刻，終話時刻から，講義動画に字幕を表示させるためのファイルを生成する．本システムの上記工程終了時，各英日翻訳結果テキストには，一行ごとにそれぞれの文の発話開始時刻，終話時刻，音声認識・修正後の英文，英日翻訳・修正後の日本語文が書き込まれている．クライアントアプリケーションが保持するすべての英日翻訳結果テキストファイルを元に，英語字幕ファイル，日本語字幕ファイルの 2 つを srt 形式で出力する．srt ファイルは，テキスト形式で記述された字幕ファイルで，動画再生時に字幕を付けることを可能とするものである．SACMI によって生成する srt 形式の英語字幕ファイルの一例を図 7 に，日本語字幕ファイルの一例を図 8 に示す．

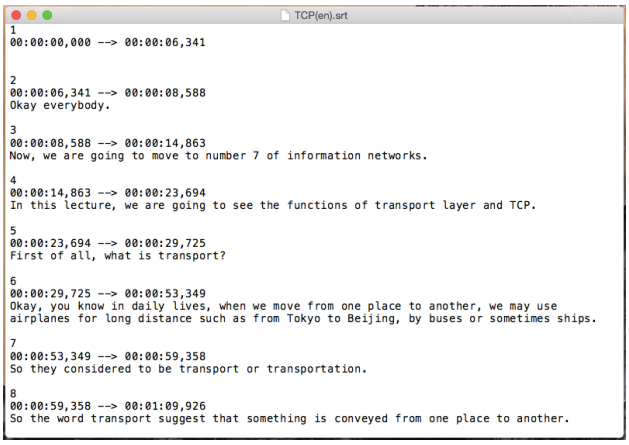

図 7 英語字幕ファイル

Figure 7 English subtitles file.

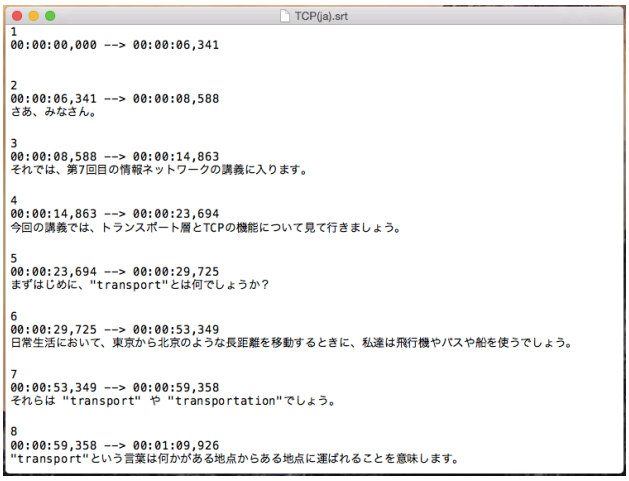

図 8 日本語字幕ファイル

Figure 8 Japanese subtitles file.

3.5 動画閲覧システム

字幕生成の工程を経て字幕ファイルを生成すると，図 9 のように，字幕付き動画を作成・閲覧することができる．そのために動画閲覧システムのサーバにアップロードする必要がある．動画閲覧システムは受講する学生がどんなネットワークや端末の環境にいても容易にアクセスできるよう Web ベースで作成した．動画の登録も Web 上から行い，サーバのデータベース上に登録していく．

データベースに登録された動画は，講義名と講演者名とともに一覧表示され，クリックすると実際に動画が再生されるページに遷移する．動画の再生と字幕の表示にはHTML5を用いて実装した．HTML5のvideoタグを用いることで動画の再生などの制御を行うことができ，videoタグ内にtrackタグを記述することでsrtファイルを読み込み動画上に数種類の字幕を選択的に表示することが可能となる．

図 9 動画閲覧システム
Figure 9 Video browsing system.

4. 評価実験

本章では，3人の被験者が実際に同一の英語での講義から日本語・英語の字幕ファイルを生成する実験の評価を述べる．実験は，SACMIシステムを用いた場合と用いなかった場合の2種類を同一被験者が行い，動画から字幕ファイルを生成するまでに要する作業時間を指標として評価を行う．

4.1 実験

本実験では，表1に示した情報工学を専攻しており，すでに当該講義を日本語版で受講済みの学生A，B，Cの3人が1分間の英語での講義(情報ネットワークにおけるTCPの初歩)を，はじめにSACMIを使用しない場合の字幕ファイル生成を行い，次にSACMIを使用する場合の字幕ファイル生成を行い，その作業時間を計測する．ただし3人ともSACMIの使用経験は全くない．SACMIを使用しない場合と使用する場合では作業工程が異なる．SACMIを使用しない場合には，被験者は音声を聞いて文章を起こし，起こした文章を翻訳し，字幕ファイルのフォーマット通りにそれぞれの文の発話開始時刻 （動画の字幕を表示させる時間） と終話時刻 （動画の字幕を消す時間）を記述する．一方SACMIを使用した場合は，動画から音声ファイルを作成し，その音声ファイルを1文ごとに切り，SACMIで音声認識・翻訳・字幕生成を行う．

表 1 被験者情報
Table 1 Details of subjects.

被験者	英語レベル	SACMI使用歴
A	TOEIC: 400	初めて
B	TOEIC: 500	初めて
C	TOEIC: 730	初めて

4.2 評価

図10のとおり，被験者AはSACMIを使うことにより，作業時間が91分から47分と48%削減することができた．同様に，被験者B・CもSACMIを使うことにより，それぞれ39%，38%削減することができた．本実験では1分間の英語による講義を，SACMIを用いて作業時間の削減を可能とした．本来の講義は60分間や90分間であるため，さらに大幅な作業時間が期待される．

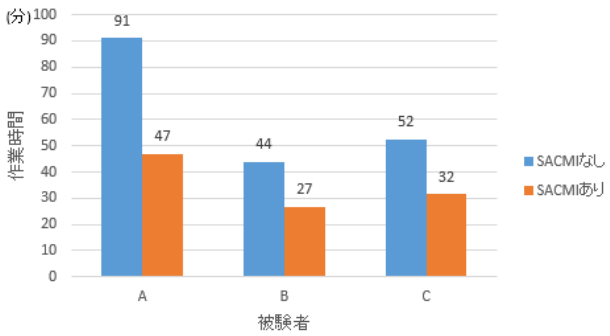

図 10 実験データ
Figure 10 Result of experiment.

5. 考察

実験結果より，SACMIを用いた方が，作業時間が短縮され，さらに出来上がった字幕ファイルもSACMIを用いた場合の方がより正確さが増していることがわかった．被験者A・B・CはSACMIを用いなかった場合，翻訳よりも音声から文章を起こすのに多くの時間を費やしていた．最初から最後まで音声を聞いて文章を起こすよりも，SACMIを用いて，完全ではなくてもある程度の英単語が用意された状態から音声を聞いて文章を修正する方が，作業時間を短縮させることができることがわかった．今回の実験では特定話者への適用のサンプル量が少なかったため，認識精度が高いとは言えなかったが，適用量を増やすことにより，認識精度も高まり，作業時間がさら

に短縮されることが期待できる．英日翻訳においても一般の翻訳エンジンでは出てこない専門用語にも対応させることができるため，英日翻訳の作業時間を短縮可能できたと考えられる．また，docomo の翻訳においても専門用語を学習させることが可能なので，サンプル数が多いほどより誤訳が少なくなると考えられる．

6. 結論

英語による専門科目の講義の字幕ファイル作成は，作業者の英語の能力と専門知識の有無に大きく依存する．音声からの文章起こしは英語の聞き取り能力や，英単語の表現の知識を要し，起こした文章は専門用語を多く含んでいるため，作業者は専門知識に精通している必要がある．そのため，どちらか片方だけでも欠けていると作業時間が大きく膨らむことや，講義を行った講師の意図していない字幕が生成される場合がある．しかし，SACMI は音声認識によってある程度の単語を認識することができ，誤認識した場合でも，発音が類似した英単語が表示されるため，編集者が英語の聞き取りが不得手でも，文章に起こすことが可能である．また，講義で頻繁に出現する専門用語は，あらかじめ辞書登録しておくことで，専門知識に精通していなくても，容易に翻訳を修正することができる．したがって，SACMI を用いることで，作業時間を削減できるだけでなく，作業者の負担も軽減することが可能である．

本研究では，英語による専門科目の講義動画への字幕作成似かかる作業時間を短くすること，認識・翻訳の精度を高めることを目的として，SACMI を開発した．今後の課題としては，音声認識に使用されている言語モデルを特定話者への適用を容易にすることがあげられる．また，英日翻訳修正画面においても，翻訳時の単語の対応関係を見られるようにすることで，専門用語の辞書登録を容易にする点に改良の余地がある．また，現段階では SACMI 使用前の事前処理として，動画ファイルから音声ファイルを人の手によって切り出しているが，その作業も SACMI の一部として自動化することで，作業時間を削減することも可能である．

謝辞

本システムを開発するに当たり，NTT ドコモサービスイノベーション部的場雄太主査（現コンシューマビジネス推進部），小野隆哉主査，清水貴司様の皆さまに技術上のご支援を賜りました．ここに御礼申し上げます．

参考文献

1) Carnegie Mellon University: Sphinx
http://cmusphinx.sourceforge.net/
2) 朝川智，峰松信明，村上隆夫，伊勢井敏子・ヤーッコラ, 広瀬啓吉: 音声の構造的表象に基づく非母語話者の英語発音分析， 電子情報通信学会技術研究報告．SP，音声 105(132)，25-30， (June 2005)
http://www.gavo.t.u-tokyo.ac.jp/~mine/paper/PDF/2005/SP2005-24_p25-30_t2005-6.pdf
3) 大崎功一，峯松信明，広瀬啓吉: 日本人英語発声に観測される発音上の癖を考慮した音声認識，電子情報通信学会技術研究報告．SP，音声 102(749)，7-12，(March 2003)
http://www.gavo.t.u-tokyo.ac.jp/~mine/paper/PDF/2003/SP2002-180_p7-12_t2003-3.pdf