実世界情報システムプロジェクト
視聴覚研究グループ峯松研究室 言語バリア研究ユニット

峯松 信明

情報理工学系研究科 電子情報学専攻

概 要

本研究ユニットは，人と人との円滑な意志疎通を，時として，妨げる「言語バリア」に対して技術的解決を検討する研究ユニットである。本ユニットは大きく 2 つに分かれる。一方は，外国語（発音）学習の支援，更には，外国語学習そのものの変革を実現すべく技術開発を行なっている。他方は聴覚障害者向けに，講演の字幕化を検討している。特に速記者を見つけることが困難な非母語話者による英語講演の字幕化を検討している。国際会議などの字幕化を考えると，母語話者の音声認識技術だけではその利用価値は甚だ低くなる。

1 はじめに

人と人との意志疎通は，計算機技術，ネットワーク技術により様々な形態のものが提案・実現され，現実社会に浸透している。しかし相手と「腹を割った」意志疎通を図る場合，人はその相手が見える，聞こえる，触れる場を選ぶ。即ち，表層的な意志疎通ではなく，真の意味で相手を理解したい，あるいは，相手に理解して欲しいと望んだ時，人はバーチャルな場ではなく，相手の現実的な存在を身近に欲求する。しかし，相手が目の前に現実的に存在していたとしても意志疎通が図れるとは限らない。その一つの理由が言語バリアである。言語バリアを考える場合，所謂外国語の問題と，聴取や調音に障害を持つ場合の 2 つが考えられる。前者に対しては外国語教育において様々な試みが行なわれてきたが，現状を顧みるとその成果は十分とは言えない。本報告書では，日本人による英語学習に対して，学習支援と言う観点のみならず，教育をどう変えるべきか，という観点からの試みを報告する。一方後者に対しては「バリアフリー」という言葉が広く使われるようになり，利用者を限定せずに，同品質のサービスを提供できる枠組みが求められている。本報告書では，日本人英語音声認識の高精度化について報告する。これは国際的な発表の字幕化以外の応用についても視野に入れている。マルチメディア DB に対する情報検索において音声認識結果が利用されるようになった現状を考慮すると，国際的な声明発表などを検索するためには，各言語を母語とする話者による英語音声認識技術を整備する必要がある。本研究はその日本人（語）版の開発に相当する。

昨今，国際的な衝突を非常によく見聞きする。その多くは，両者の意志疎通，共通理解の不足に根づく部分も多い。様々な意志疎通メディアが導入されたにも拘らず，真の意味での理解がまだまだ不足していることが示唆される。本研究ユニットの究極の目標は，人と人が「本当に」伝えたいことを，両者が「深く」理解できる言語メディアで,「腹を割って」共感することができる「世界」の創出である。そのための技術支援を目指している。

2 英語学習の支援・変革を目的とした研究開発

中学に入学する時，誰もが「英語を使って意志疎通できる」自分を夢に見る。そして凡そ 8 年に渡って英語教育を受ける。しかしその夢を実現できる

人間は非常に少ない。事実，TOEICによる国別点数によれば，日本人学生の点数はアジア諸国の中で最下位にある。この状況を打破すべく，音声情報処理技術を用いた種々の試み（CALL=Computer Aided Language Learning）が行なわれている。しかしその多くは，バーチャルな教師をPC上に登場させる，という量的解決を図ることを目的としている。（世界的に見れば極めて）勤勉な国民性を持つ日本人学生が8年もの長い年月をかけて最下位の能力しか獲得できない現実を考えた場合，それは量的解決ではなく，教育の質的改革の必要性を示唆しているように感じる。以上を考慮し，本研究では，従来とは観点の異なる発音教育支援を目的とした技術開発を行なった。

2.1 英語発声におけるエチケットの追及

2.1.1 日本語の「耳」と英語の「耳」

この文書を読んでいる読者のうち，自らが発声した英語を実際に母語話者に書き取ってもらい，自らの英語のどこが「聞き取れない，書き取れない」のかを調査した読者はいるだろうか？留学経験が無ければ，間違いなく皆無であろう。日本人と英語母語話者の音声知覚プロセスの違いがしばしば議論される。音の取り入れ方，音声連続体の区切り方が根本的に異なるのである。音声認識装置を実際にユーザに使わせると，誤認識された場合，機械に伝わるよう極端に「ゆっくり，丁寧に」喋り出す。しかし，そのような音声は機械にとっては「不明瞭」極まりない。機械の「知覚」は学習用音声データが全てであり,「ゆっくり，丁寧な」発声は通常学習データに存在しなため，結局，極めて「不明瞭」は発音となる。同様のことが，日本人と英語母語話者間で起きる。「明瞭にしよう」という努力がかえって「不明瞭」な発音を生む。話し手と聞き手とで異なる「耳」を持っている，と言うこともできる。一方海外に留学すると，毎日自分の英語を「聞き取って」もらい，分らなければ「Sorry ?」と問い返される。つまり，相手の「耳」の特性を24時間学習することになる。その結果「聞いてもらえる」英語は最低限身に付く。

2.1.2 Virtual Native Earsの構築

「聞いてもらえる英語」を習得する場合，幾ら英語を読んでも，書いても，聞いても,（独話的に）話しても，効果は薄い。「何故相手が聞き誤るのか」に対する解答は，相手不在の環境では（直接的には）得られないからである。そこで本研究では，まず幅広い発音習熟度を持った日本人学生男女100人ずつの読み上げ英語音声を収録した。その中から文長，言語的複雑さ，発音能力についてバランスをとった360文発声セットを定義し，これを実際にアメリカ人，カナダ人に聴取させ，書き取らせた。「何がいけないのか」を客観的に，具体的に議論するための基礎データの収集である。

次に，この文セットを音響分析，言語分析にかけ，種々の音響属性，言語属性の値を抽出した。音響属性に関しては，音韻性を伝搬する分節的特徴や，英語リズムを伝搬する強弱勢に関する属性などを定義し，その値を単語（一部は文）を単位として抽出した。言語属性としては，品詞，単語数やN-gram値など，広く自然言語処理で利用される言語属性を定義し，その値を抽出した。

得られた単語単位（一部は文）の属性値を説明変数として，CART分析により，英語母語話者による書取り率を単語単位で予測することを試みた。つまり，日本人英語の音的要因，言的要因の何と何が絡んで「聞き取れない」という現象を生むのか，を純粋にボトムアップ的に検討した。

また，日本人の「耳」を持つ日本人の英語教師にも同様のことを依頼した。書取り率が0/6〜6/6（被験者数6人）の7段階で定義されているので，文音声の聴取後，各単語に対して7段階で「聞き取ってもらえる可能性」を評価してもらった。

結果を表1に示す。±1の誤りを無視したrecall, precisionをチャンスレベル（C），日本人英語教師（1名，H），計算機（M）として示している。まず，英語教師による予測であるが，4/6以下のrecallがチャンスレベルを大きく下回る。本実験は「聞き取れるかどうか」を議論しているが，これは人間の高度な適応能力を議論していることになる。結局，音声の知覚プロセスとして母語話者と等しいものを保持してことが要求されるタスク

表 1: 予測正解率 [%]

	0/6	1/6	2/6	3/6	4/6	5/6	6/6
C recall	28.5	35.4	43.3	43.8	42.9	43.5	29.8
prec.	7.1	11.0	15.6	22.6	33.0	79.4	73.4
H recall	2.5	5.8	6.8	13.3	10.1	95.8	93.6
prec.	12.5	40.9	24.1	41.8	24.7	74.2	73.8
M recall	67.8	85.7	84.2	75.0	71.7	75.9	59.7
prec.	38.2	46.1	28.3	44.2	50.0	95.8	93.6

であり，上記の結果が出たと考察している。一方機械による予測であるが，特に正解率が低いものに対して高い recall を呈しており，その有効性が十分に示されている。今後はデータを増やすと共に，新たな説明変数の導入などを検討している。

なお，平均書取り率は，日本在住の母語話者で約 80%，日本人と会話したことが無い母語話者で約 70%である。これは，ノイズレベルに換算すると，約 −1.2dB，−3.3dB の SN 比に相当する。

2.2 分節的な明瞭度の高い英語の必要条件

2.2.1 日本語の「口」と英語の「口」

従来の発音教育は音声学の知見に根差すものが多い。音声学は，個々の音（単音）がどのような調音運動によって生成されるのかを議論する学問であり，個々の音の正しい発音を目的とした場合，有益な情報源となる。しかし「調音的に誤った発音が誤聴取を導くのか」という問いに答えることはできない。音声学の基本は音声生成であり，音声知覚ではない。さて，正しい調音は日本人の口で実現できるだろうか?日本語に慣れ親しんだ「口」には英語を正しく調音するための筋肉が十分には備わっていない，という議論を考えると，舌の位置，口唇の形を図示されても，それを実現する土台が無い日本人の「口」では困難である。右利きの人に，左手で文字を書かせるようなものである。奇麗な文字の形（正しい舌の位置，口唇の形）が示されても，左手での再現は困難である。英語教育者の中には，呼吸法も含め，肉体訓練（改造）から教育を開始する者もいるが，少数である。

2.3 「母語話者発音」に対する必要条件

「調音的に正しい発音」を「母語話者発音」の十分条件とした場合，必要条件は何であろうか？十分条件の充足が困難であれば，必要条件を考え，それを目標とするのも現実的な方法論である。

米語話者音声データから米語音響モデルを，日本人英語音声データから日本人英語音響モデル（HMM，各音素は 3 つの状態の left-to-right な遷移で表現される）を構築した。次に任意の 2 状態間の距離を算出し，状態間距離行列を求めた。この距離行列に対して階層的クラスタリングを施すことにより状態を単位とした樹型図が描かれる。誌面の都合で，日本人英語に対する樹型図だけを示す（図 1）が，母音混同，子音混同，母音挿入など従来知られた日本人の発音の癖が克明に反映されている。この樹型図は距離行列より形成される。距離行列は，音素モデルセットから各音素間（状態間）の距離だけを抽出し，各音素の音響空間内における絶対的位置の情報は切り捨てている。音響空間内の位置は「音声のスペクトル包絡情報」に相当し，これを音声学の言葉に置き換えると「調音位置，調音形態の情報」となる。つまり，どのような調音運動によって日本人が英語音声を生成したのか，という情報を全く捨てた分析であるにも拘らず，従来広く指摘されている日本人の発音上の癖は明確に残っている。

本研究では「個々の音を調音的に正しく生成すること」を十分条件とした時に,「発音全体において各音が正しい関係で配置していること」を必要条件とし，後者を目指す教育論の可能性を検討している。現在，各学習者の木を個別に生成し，DB 中の習熟度ラベルとの関係を分析している。

3 日本人英語音声認識の高精度化

現在の連続音声認識技術は，音響モデル，言語モデル，発音辞書，デコーダーとモジュール化されており，非母国語の音声認識を検討する場合も，何れかのモジュールの改良，ということになる。ここでは音響モデルの高精度化を検討した。

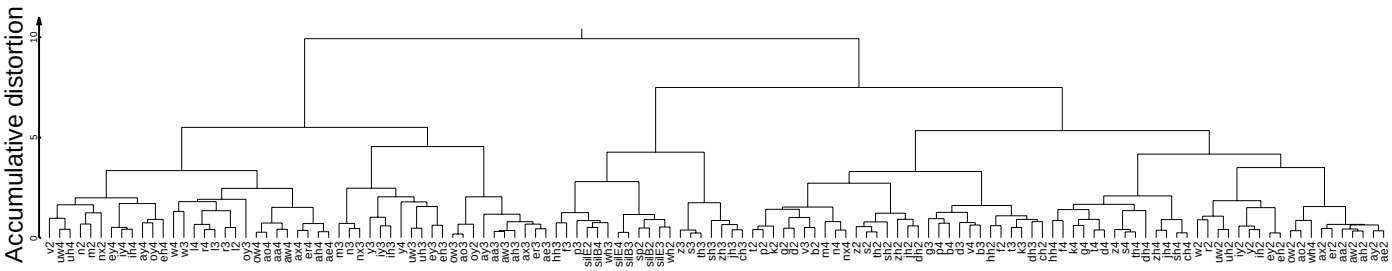

図 1: 日本人英語から構築された音素状態を単位とする樹型図（ウォード法使用）

3.1 スペル情報による発音歪みのモデル化

triphone モデルを構築する場合，パラメータ数の増加を抑えるために状態共有が行なわれる。この場合，音声学的な発音構造の知識をトップダウン的に与え，状態分割決定木を構成する。しかし，日本人英語には英語音声学が規定する以外の特性が観測される。その一つがスペルによる発音歪みである。本研究では英語母音をスペル情報を付与すことで拡張し（母音数は 16 から 95 へ増加），拡張母音を用いた triphone を効率良く使用することで認識率の向上を実現した。

3.2 習熟度に応じた発音構造を用いた適応

日本人英語不特定話者モデルに対して話者適応を行なうことで認識率の向上は実現できるが，この場合，適応においても発声に内在する音声学的構造を前提とした処理が行なわれる。しかし，日本人英語は幅広い習熟度のために，その構造を事前に予測することが困難である。ここでは，適応データから当該話者の発音に見られる音声学的構造を推定し，その構造に基づいて適応をかけることで認識率の向上を実現した。

3.3 日本語モデルの併用による高精度化

発音習熟度の低い話者に着眼すると，日本人英語モデルよりも日本人日本語モデルを用いた方が尤度が高くなる発声が頻繁に見られるようになる。そこで，日本人英語 triphone と日本人日本語 triphone の対応を音響的に定義し，これらを並列モデルとして実装することで認識率の向上を実現した。並列化のためには，分岐確率の適切な推定が必要となるが，ここでは，適応データに対する最尤パラメータを求める形で算出した。

4 まとめ

「言語バリア」の技術的解決に関する検討を報告した。誠に微力ではあるが，言葉の壁，言語能力の壁を越え，人と人とが真の意味で対話できる「世界」の実現に向けて尽力していきたい。

発表文献（一部）

[1] N. Minematsu *et al.*, “English speech database read by Japanese learners for CALL system development,” Proc. LREC, pp.896–903 (2002)

[2] N. Minematsu *et al.*, “Acoustic Modeling of Sentence Stress Using Differential Features between Syllables for English Rhythm Learning System Development,” Proc. ICSLP, pp.745–748 (2002).

[3] C. Guo *et al.*, “Prediction of American listeners’ misrecognition of English words spoken by Japanese”, Technical report of IEICE, SP2002-179, pp.1–6 (2003)

[4] N. Minematsu *et al.*, Corpus-based Analysis of English spoken by Japanese students in view of the entire phonemic system of English, Proc. ICSLP, pp.1213-1216 (2002).

[5] N. Minematsu *et al.*, “Corpus-based Analysis of production and perception of Japanese English in view of the entire phonemic system of English,” Proc. ICPhS, (2003, to appear)

[6] N. Minematsu *et al.*, “Integration of MLLR Adaptation with Pronunciation Proficiency Adaptation for Non-native Speech Recognition”, Proc. ICSLP, pp.529-532 (2002).