

論文 / 著書情報  
Article / Book Information

|                   |                                                                                                                      |
|-------------------|----------------------------------------------------------------------------------------------------------------------|
| 論題(和文)            | 放送ニュースの認識と理解に関するDARPAワークショップ                                                                                         |
| Title(English)    |                                                                                                                      |
| 著者(和文)            | 古井貞熙                                                                                                                 |
| Authors(English)  | SADAOKI FURUI                                                                                                        |
| 掲載誌(和文)           | 情報・システムソサエティ誌                                                                                                        |
| Citation(English) |                                                                                                                      |
| Vol, no, pages    | Vol. 3, No. 2, pp. 20-21                                                                                             |
| 発行日 / Pub. date   | 1997,                                                                                                                |
| URL               | <a href="http://search.ieice.org/">http://search.ieice.org/</a>                                                      |
| 権利情報 / Copyright  | 本著作物の著作権は電子情報通信学会に帰属します。<br>Copyright (c) 1997 Institute of Electronics, Information and<br>Communication Engineers. |

## 放送ニュースの認識と理解に関するDARPAワークショップ

古井 貞熙(東京工業大学)



米国のDARPAの音声認識プロジェクトは、その具体的内容を数年毎に変えているが、20年近くにわたって、世界の音声認識研究の主要な牽引車であり続けてきた。本国際会議("DARPA Broadcast News Transcription and Understanding Workshop")は、そのプロジェクトが毎年春に行っているワークショップの一環である。会議の名称はその年の中心的課題によって変化している。今年の会議は、昨年から活発になった放送ニュース音声のディクテーション(Hub4タスクと呼ばれる)を中心的話題として、ワシントン郊外のLansdowne Conference Resort(Virginia州)で、2月8日から11日まで行なわれた。参加者は主催者からの招待者のみで、約200名であった。会場は、米国の典型的なリゾートで、広大なゴルフ場の中心にある豪華なホテルで、合宿形式で行なわれた。



Hub4タスクでは、種々の放送ニュース番組を収録した大量の音声データベース(のべ約100時間)と、それを文字で書き起こしたものを基本とする大量のテキストデータベース(1~5億語)が用意されている。それらを共通に用いたコンテストの形で、米国とヨー

ロッパの多数の研究機関が参加して、研究が行なわれている。基本的なアルゴリズムはほとんどの研究機関に共通で、HMMによる音素環境依存形音素モデルと、トライグラムによる統計的言語モデルが用いられている。語彙数は65000語が普通である。あらかじめ原稿が用意されているアナウンサーの音声から電話やインタビューまで、対象の難しさが、表1に示すF0からFXまでの7種類の条件(focus condition)に分類されている。昨年は各条件のテスト音声の量がバランスされていたが、今年は、実際の放送における出現頻度に対応した量になった。認識性能を上げるため、これらの種類に応じて、異なる音素モデルが使われる。昨年の評価実験では、評価用音声は、マニュアルで7種類の条件にあらかじめ分類された実験(PE; Partitioned Evaluation)と、分類されていない実験(UE; Unpartitioned Evaluation)の2種類が行なわれたが、今年はUEだけになった。従って、入力音声を自動的にいずれかの条件に分類することが必要になるが、このプログラムは、各研究機関が自分で作るか、CMUから提供されるものを使っている。

分類された各条件の音素モデルは、各話者の声の違いに対処するために、不特定話者用のモデルに対して、MLLR(Maximum Likelihood Linear Regression)法などを用いて、オンライン教師なし話者適応を行うのが普通である。あらかじめ多数話者の音声を用いて不特定話者音素モデルを作る際に、話者適応を行いながら作成するSAT(Speaker Adapted Training)と呼ばれる方法を用いている機関もある。いずれの研究機関でも、昨年のシステムと比べて特に大きな技術的進歩があるわけではないが、種々の技術的改良と着実な技術の積み上げによって、認識性能

表 Hub4における実験条件

|    |                                                                 |
|----|-----------------------------------------------------------------|
| F0 | ベースライン放送音声 (スタジオで録音した原稿のある音声)                                   |
| F1 | 対話放送音声 (スタジオで録音した対話音声)                                          |
| F2 | 電話系音声 (帯域制限された音声)                                               |
| F3 | 背景音楽中の音声 (SNRがA特性で10~20 dBの原稿あり/なし音声)                           |
| F4 | 劣化音響条件下の音声 (SNRがA特性で10~20 dBの、加算性雑音、環境雑音、あるいは非線形歪みのある原稿あり/なし音声) |
| F5 | non-native話者の音声 (米語のnon-native話者—英国のnative話者の英語を含む—のスタジオ品質の音声)  |
| F6 | その他 (上に属さないものや、上の2つ以上の条件を満たすもの—例えば、背景音楽中のnon-native話者の音声—)      |

が向上している。最も基本的なF0の条件で、主要な研究機関の単語認識誤り率は昨年の約1/2に低下しており、最も高い性能を示したシステムで、約90%の平均単語認識率が得られている。F0からFXまでの平均では、約85%の平均単語認識率が得られている。

実際に使えるシステムを作るためには、認識処理速度の向上が必須であるが、そのためには、今のところいくつかの機能を省略することが必要であり、そうするとどうしても認識性能の低下が避けられない。実時間の約200倍の時間がかかるシステムの単語認識誤り率が20.3%であるとき、実時間の6倍に高速化したシステムでは25.7%に増大するという結果が報告されている。

今年の特徴としては、放送以外の演説 (大統領選挙の演説) や、英語以外の言語として、スペイン語と中国語の放送ニュースが、研究対象に含まれるようになったことがある。演説の音声に対しては、上記のF0の条件の約1.5倍の単語認識誤り率が得られている。スペイン語と中国語の放送ニュースに対しては、英語とほとんど同じ認識率が得られると報告されているが、まだ参加している研究機関やデータ量が少ないので、今後さらに評価実験を進める必要がある。日本語の放送ニュースディクテーションに関しては古井が報告したが、DARPAのプロジェクトで日本語を取り上げる計画はない。

研究の流れとしての今年の特徴としては、音声認識技術と情報検索技術の結びつきにより、

音声からの自動的な情報抽出 (IE; Information Extraction), 情報検索 (IR; Information Retrieval), トピックス検出と追跡 (TDT; Topic Detection & Tracking) などの研究が、極めて活発に行なわれるようになったことがある。基本的な方法としては、音声認識によって音声を単語列に変換し、その単語列から、トピックス (話題) を表す単語を自動的に抽出する。これにより、音声データにインデックスを自動的につけることができ、音声データの情報検索 (SDR; Spoken Document Retrieval) の手掛かりとなる。音声から、その意味に対応する言葉を自動的に抽出する研究として、人名、地名、組織名などを表す言葉を抽出する試みも始まっている。

DARPAの音声認識プロジェクトの新しい研究目標として、ニュース音声のディクテーションとは別に、音声対話による情報検索システム ("DARPA communicator") の研究が提案された。以前にDARPAプロジェクトとして行なわれたATIS (航空旅行情報システム) タスクの再現と見ることもできる。来年に向けて、このような研究も活発になってくると思われる。

本ワークショップの最終的な論文集は、例年通り、数か月後に出版される予定であるが、今年は電子的な形で出版も行なわれる予定である。