大学の講義映像に対する音声認識技術を用いた字幕制作支援

A support of subtitling with a speech recognition technology for a video of university lectures

松田　祥
Sho Matsuda
法政大学情報科学部ディジタルメディア学科
E-mail: sho.matsuda.zy@cis.hosei.ac.jp

Abstract

A support of information by sign-language interpreter and Note-take, and more is necessary for understanding contents of a lecture for students with hearing impairment who learn at a university. And it will be able to become useful support of information for not only a hearing impairment student but also hearing people if there is a lecture video with subtitles. Purpose of this study is to support creating the subtitles of the lecture video with speech recognition system. This study directly recognized without a repetition of third party a speech of recorded lecture. A lecture includes original expressions and technical terms to each subject, and words not registered in the word dictionary in a speech recognition system used cannot be recognized. Then, this study chose the technical term not registered from lecturer's material, and registered their words to the word dictionary in the speech recognition system. As a result, the recognition rate before it registered was 33%, and the recognition rate after it had registered was 36%. There were data with a lot of numbers of dropouts of the word in result of recognition. Then, this study changed a value of the word insertion penalty for only their words, and their data was recognized again. As a result, the recognition rate became to 37% from 36%.

1. まえがき

現在日本の大学において聴覚障害学生に対する情報保障は十分とは言えない．そのため健聴者が講義で受け得る情報量に匹敵する程の情報を聴覚障害者に与える必要性は極めて高い．また，ブロードバンドの普及により，近い将来，大学講義の映像がインターネット上で配信される可能性も十分ある．そういったサービスに対しても字幕の付与は大切と言える．音声認識技術を用いた講義音声に対して講義内容の要約を検討する研究[2]もあり，音声認識技術による自動化への期待は大きい．

群馬大学では音声認識システムを用いたノートテークによる学生支援が試験的に行われている[1]．また，愛媛大学ではインターネットを介し，遠隔地から音声認識システムを用いた情報保障システムが開発されている[3]．

しかし，いずれの場合もリスピークと呼ばれる方式を用いていて，これは発話者とは別の人物が丁寧に復唱したものを音声認識システムに入力している．この方法を用いると認識精度は高くなるが，復唱者は訓練をしなければならない上に，1 人で復唱し続けるわけにもいかず，最低でも 2 人交代で行う必要がある．また，誤った認識を修正する者も必要になる．つまり，最低でも 3 人の人間が必要という事になる．これは人員の確保やシステムの環境作りが問題になってくる．

本研究では大学講義に対するリアルタイムの字幕付与ではなく，音声認識技術を用いて，ビデオ撮影した講義映像に対する字幕制作の支援を目指す．

字幕を制作するに当っては講義内容，つまりは講師の発した言葉を文字に書き起こす必要がある．講義は約 90 分間あるため，その間の発話数は相当な数になり，書き起こしは大変な作業である．そこで講義映像の音声データを直接認識させる事が出来れば，修正者 1 人だけで済むので，字幕制作のための書き起こし作業の負担が大幅に軽減されると考えられる．

また，素人によるリスピークで良好な結果が得られるのであれば書き起こし支援という点においては有効だという可能性もある．

2. 音声認識技術における問題点

現在の音声認識技術では自然な会話，つまり口語的な表現が多い場合や，早口やささやき声といったような場合には認識は非常に困難になる．本研究で取り扱う大学講義はまさにこういった問題に直面する．講師により差はあるものの多くの講師が口語的な表現で発話しているし，話が盛り上がってくると早口になる傾向もある．また，ドラマなどと違って台詞が決まっている訳でもないので，文頭や文中に「えー，まー，そのー」などの言い淀み(フィラー)が多い事も特徴として挙げられる．

このような問題を回避するためには，技術的な限界を考えると，講師がある程度ゆっくり話す事や言葉遣いに注意を払うといった講師側の協力も必要になってくる．しかし，一定速度で長時間話すにはそれなりに訓練が必要であり，講師全員に訓練を強要するのは現実的に困難である．講義には専門用語が多く含まれていて，その未知語によって認識率が低下していると考えられる．また，発話単語数に対しての極端な単語脱落も認識率低下の原因の一つと考えられる．そこで，未知語に対しては単語辞書の登録で，単語脱落に対しては単語間の遷移の抑制，促進を変動させ，認識結果に直接的に影響を与えるパラメータである「単語挿入ペナルティ」の設定などを変更する方法で認識率の向上を目指した．

また，リスピーク方式で音声認識させて字幕を制作する場合，訓練の必要性や人材確保などの問題があるが，素人によるリスピークで認識率がどの程度になるかを検証した．

Supervisor: Prof. Katsunobu Ito

3. 講義音声認識実験

講義音声データを直接入力し，音声認識システムがどの程度認識できるのか実験した．辞書登録前と後とでの認識率の比較を行った．

3.1 実験用音声データ

音声データは，2008 年 7 月 8 日に行われた狩野教授の「微積分学」，同年 11 月 21 日に行われた佐々木教授の「離散数学」の二つの講義を撮影した映像データの音声を使用する．どちらの講義も学部 1 年生向けで，講義時間は約 90 分である．この二つの講義内容の書き起こしを実際に行った結果を以下に示す．狩野教授，佐々木教授の講義をそれぞれ A，B とした．

	書き起こし作業時間	書き起こし文字数	含まれるフィラーの数
講義 A	11 時間 43 分	約 25,000 字	266 回
講義 B	7 時間 40 分	約 12,500 字	217 回

3.1 .1 収録方法

撮影は，教室の一番後ろの中央にビデオカメラをセッティングし，音声は講師の胸元あたりにヒモでぶら下げた Bluetooth 対応(カメラとセット)のワイヤレスマイクで録音した．なお，48kHz，16bit で録音．カメラは図 1，マイクは図 2．撮影全体の様子は図 3，マイク装着の様子は図 4 を参照．

図 1　HDR-CX7(SONY)

図 2　ECM-HW1(SONY)

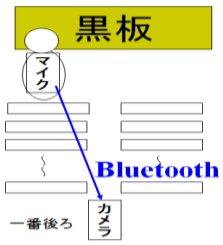

図 3　撮影風景

図 4　マイク装着の様子

3.1 .2 音声データ切り分け

また，音声認識システムに入力するにあたって音声データを下記の 2 つの条件によって切り分ける．

・ひとつのデータは約 3 秒から 10 秒

・単語数は 5 単語から 45 単語程度

認識結果との比較を行うため，この条件によって切り分けた音声の書き起こしを行った．音声は切り分けた順にデータ番号を付けた．その例を表 1 に示す．

これらの音声データを Julius で認識させるために，講義映像の音声を WAV 形式の 16kHz，16bit，モノラルに変換する処理を行った．

表 1　講義音声の書き起こし例

番号	内容
001	さぁそれじゃあ授業をスタートいたしましょう
002	ええ，今日含めてあと残りが 2 回ですか
003	ええ，もう極わずかで学期末試験ですね

3.2 音声認識システム Julius

音声認識には大語彙連続音声認識エンジン Julius [1]を使用する．Julius は与えられた音響モデルや言語モデルを用いて認識するものである．よってこれらのモデルを変える事でさまざまな入力環境やタスクに適用できる．

本研究では Julius ディクテーション実行キット v3.2 を使用する．音響モデルの変更や単語辞書の追加などによって認識結果がどう変わるかを検証していく．

使用した音響モデルは「JNAS 標準成人モデル」で，これは「新聞記事読み上げ音声コーパス」により学習したモデルで，IPA「日本語ディクテーション基本ソフトウェア」で使用されるモデルと同じものである．

Julius では単語辞書に認識対象の語彙とその発音を規定していて，これに規定されているもののみがマッチングの対象となる．つまり，この辞書に登録されていない単語は一切認識できない．

3.3 講義用資料

一般に，講義を行うに当って講義用の資料を用意してある事が大体である．その例としては，教科書，配布用プリント，講師自身用講義資料，スライドデータなどが挙げられる．これらの資料から専門用語や特別な言い回しなどを参考にし，未知語として辞書に登録する．

3.3.1 辞書登録に使用する資料

大学講義では専門用語が多く登場するため，科目ごとに単語辞書を用意する必要がある．

今回実験で使用する講義 A には講師が自身用に用意した当日に行う講義内容の資料が A4 サイズ，両面印刷，計 4 枚あり，講義 B には講師自筆のノートが 2 枚と演習プリントが 1 枚あった．教科書を参考にするよりは，ピンポイントでその日の内容が書いてあると思われたので，この資料を参考にして辞書に単語を追加する事にした．

3.3.2 単語辞書登録

ディクテーションキットのデフォルトの単語辞書に，先に述べた講義用資料を参考にして講義 A，B ともに 20 単語を追加した．

表 2　辞書登録した単語

講義	登録単語
A	偏微分/ 偏導関数/ 数Ⅲ/ 学期末/ 授業/ df/ dx/ 偏微導関数/ 陰関数/ 陽関数/ 極値/ 接平面/ 高階偏導関数/ 鞍点/ 定義域/接線/ ラグランジュ/ 助変数/ 全微分/ 極座標/
B	直積/ 部分集合/ 順序対/ 座標図/ 関係行列/　関係グラフ/ 節点/ 有向辺/ 有向グラフ/逆関係/　逆行列/ 転置行列/ 恒等関係/ 推移的/ 反対称的/ 対称的/ 閉包/ 同値関係/ 中への関係　　/上の関係

[1]大語彙連続音声認識エンジン Julius
http://julius.sourceforge.jp/

辞書登録には区分，表示単語，音素表記の 3 つの要素が必要で，今回の登録において区分は全て「未知語」とした．登録した単語の詳細を表 2 に示す．

3.4 認識率

認識率を求める評価尺度として、単語正解精度を用いた．単語正解精度の定義を以下に示す．

単語正解精度＝(N – D – S – I) / N

N:正解の単語総数，D:脱落誤りの数，S:置換誤りの数
I:挿入誤りの数，C:正解

以下に計算例を示す．

対応	S	S	C	C	C
正解	微分	記号	を	この	丸
結果	自分	規模	を	この	まる

対応	C	S	I	S	S
正解	で	d		と	書いた
結果	で	い	人	とか	言った

N = 9, D = 0 , S = 5 , I = 1 , C = 4
単語正解精度 = (9 – 0 – 5 – 1) / 9 = 3 / 9 ≒0.33

4. 実験結果

4.1. 辞書登録前後による認識結果の比較

切り分けた音声データ 001 から 100 までの 100 個を Julius によって認識させ，辞書登録前と後とでの認識率を計算した．その結果を以下に示す．また，辞書登録前と後での認識結果例を表 3 に示す．表 4 に登録前後の登録単語の出現回数と正解数の比較を示す．

・辞書登録前

	N	D	S	I	C	認識率
講義 A	1919	431	791	73	689	33%
講義 B	1558	382	657	42	519	31%

・辞書登録後

	N	D	S	I	C	認識率
講義 A	1919	437	747	46	735	36%
講義 B	1558	368	608	32	582	35%

表 3　辞書登録前後の結果比較例(講義 A)

	正解	認識結果
登録前	で，えぇ偏導関数、偏微分のお話しをしておりました	で山道が数編別の話を売っておりました
登録後		で，偏導関数偏微分の話を売っておりました．

表 4　登録単語の出現回数と正解数の登録前後結果比較

	出現回数	正解数 (登録前)	正解数 (登録後)
講義 A	52	3	30
講義 B	56	8	47

4.2. 100 個のデータの認識結果に対する考察

辞書登録によって認識結果が講義 A は 3%，講義 B は 4%上がった．正解単語数 C が大幅に増加した事と，出現回数に対する正解数の増加した事から，辞書に登録されていない事だけが理由での誤認識が多い事がわかった．

文頭に多い「で」や「ええ」などの話し言葉特有の現象であるフィラーによって，誤った認識が多くみられたように思う．こういった書き言葉と話し言葉の違いにどう対処していくかを考える必要がある．また，脱落誤りが多い事から Julius の「挿入ペナルティ」のパラメータが強すぎる可能性もあると考えられる．

5. リスピークしたデータを認識

1 章で述べたように，リスピークによる音声認識は認識率が高い．例えば，NHK においてはスポーツ生中継番組に対するリスピーク方式による音声認識で 90%を上回る認識率を達成した[4].

上記の NHK の実験ではリスピークをアナウンサーが行っている．そこで，本研究ではアナウンサーのように訓練されている者ではなく，素人がリスピークをしても効果が見られるかを実験するため，講義 A，B のそれぞれ 100 個のデータのうちで，認識率が低かったもの，0，10，20，30，40，50%の 6 段階のデータをそれぞれ一つずつ，計 12 個のデータを選び認識させた．

なお，それぞれ 10 発話ずつ，計 120 発話のリスピークデータを使用した．録音は講義を撮影した時と同じ環境で，リスピークは自分自身で行った．

5.1. リスピーク実験結果

元データとリスピークデータの認識率とそれぞれのモーラ速度の平均値の比較を表 5 に示す．リスピークは元のデータの言葉を変えずに，ゆっくりはっきりと発音して行った．なお，モーラ速度とは，一つの発話に含まれるモーラの数を秒数で割ったものと定義し，単位は「mora / s」．

モーラとは，音韻論上、一定の時間的長さをもった音の分節単位で，日本語の場合は仮名ひとつが 1 モーラにあたる．拗音は 1 モーラとならず，「きゃ」の場合はこれで 1 モーラとなる．長音「ー」，促音「っ」，撥音「ん」は 1 モーラとされている．

表 5 元データとリスピークデータの認識率比較

講義	元データ認識率 (平均)	リスピーク認識率 (平均)	元データモーラ速度 (平均)	リスピークモーラ速度 (平均)
A	26%	85%	8.1	5.9
B	25%	91%	7.2	4.3

5.2. リスピーク実験結果の考察

結果，素人がリスピークしても認識率は高くなった．平均の認識率で比べると，元データは 26%，リスピークデータは 88%となった．この事から，元のデータの認識率が低かったのは言語モデル自体に原因があるのではなく，発声や話す速度による影響が大きいと考えられる．また，リスピークでの認識率が元の認識率にあまり影響されない事もわかった．この結果から，リスピークによる支援も有効な手段の一つとして考えられる．

リスピークは何故認識率が高くなるのかという事を認識率と平均モーラ速度，認識率とモーラ長の分散という 2 項目に関して調べたが，相関はほとんどなく，理由はわからなかった．

6. 挿入ペナルティ値変更

4.1 節の認識結果より，脱落誤りの数が多い事がわかった．その理由の一つとして、Julius の挿入ペナルティが強くかかり過ぎている事が考えられ，それを検証するために挿入ペナルティの数値を変更して認識した．

言語重みと挿入ペナルティのデフォルト値と変更後の値を以下に示す．lmp，lmp2 はそれぞれ第 1 パス，第 2 パスにおける言語重み(左の値)，挿入ペナルティ(右の値)のパラメータとなっている．

デフォルト値	変更後の値
lmp 8.0 -2.0	lmp 8.0 3.0
lmp2 8.0 -2.0	lmp2 8.0 3.0

6.1 実験結果

ペナルティの値を変更後，脱落誤りが減ったデータ例を図 5 に示す．変更後に挿入誤りが増加し，認識率が低下した例もあった．図 5 は上が変更前，下が変更後の結果で、不正解単語には下線を引いた．この結果から，デフォルト値で全てのデータを認識させ，脱落の多かったデータのみ，ペナルティ値を変更した設定で認識させる方法をとった．脱落の多かったデータとは，正解の単語総数に対しての脱落誤りの数の割合が 30%以上のものとした．この方法で認識させた結果を表 6 に示す．

```
こんな 風   な 話 も 習っ     た わけ です ね
この アホ な 話 も 納得 し た

これ が 全微分 という もの    の 復習 で あり ます
これ が 全微分 注目   する
                ↓
こんな 風   な 話 も 習っ     た わけ です ね
この アホ な 話 も 納得 し た が

これ が   全微分 という もの    の 復習 で あり ます
これ では 全微分  修   者     の 復讐 で ある
```

図 5 脱落誤りが減った例

6.2. 考察

脱落の多いデータのみに対して挿入ペナルティ値を変更させ認識させてみたが，講義 A，B ともに認識率の結果に目立った改善はみられなかった．脱落誤りの数 D や正解数 C には改善がみられた．しかし，D が減る事に比例して置換誤りの数 S が増え，ペナルティ値変更の影響で挿入誤り I の数が増える．これらの原因によって結果的に認識率にはあまり動きが見られなかったと考えられる．

表 6　単語挿入ペナルティ変更前後の認識率比較

・講義 A

	N	D	S	I	C	認識率
変更前	1919	437	747	46	735	36%
変更後	1919	368	774	64	735	37%

・講義 B

	N	D	S	I	C	認識率
変更前	1558	368	608	32	582	35%
変更後	1558	299	651	52	608	36%

7. むすび

本研究では，Julius による講義音声の直接認識を試み，単語辞書登録による効果，単語挿入ペナルティ値変更による認識結果の変化を実験した．また，素人によるリスピークの効果についても検証した．

その結果，単語辞書の登録と単語挿入ペナルティ値変更だけでは，認識率に大きな改善は見られなかった．

素人によるリスピークでも高い認識率が得られた事から，講義の字幕制作支援をする上で，有効な手段の一つとして考えられる事がわかった．これらの結果から，講師自身がリスピークを行っても高い認識率が期待でき，現状ではリスピーク方式が書き起こし支援には合っていると言える．

また，全体的に言える事としては，文頭，文中に多く出現する，「で」や「ええ」などの話し言葉特有の現象であるフィラーによって，誤った認識が多くみられたように思う．こういった書き言葉と話し言葉の違いにどう対処していくかを今後考える必要がある．

講義音声を認識させるにあたって直面する問題点の一つとして，英字の大文字と小文字の区別など，理数系科目にとって極めて重要度の高い情報が音声情報だけではカバーしきれない点が挙げられる．こういった問題については映像情報と組み合わせて処理を行うなどする必要があるかもしれない．

謝辞

講義の撮影にご協力していただいた狩野教授，佐々木教授に感謝いたします．

文　献

[1] 金澤貴之,”大学における聴覚障害学生への情報保障の取り組み”,『聴覚障害者のための字幕付与技術』シンポジウム,pp.3-8,Oct.2008.

[2] 富樫他,“講義音声の認識・要約・インデックス化の検討(要約・分割)”, 情処学研報.SLP，音声言語情報処理，IPSJ SIG Notes, Vol.2006, No.73, pp.57-62 ,Jul.2006.

[3] 立入他,”音声認識を利用した利用した聴覚障害学生学習保障システムについて”, 信学技報.TL,思考と言語,Vol.103,No.114(20030606),pp.43-48,Jun.2003

[4] 本間他,”生字幕放送のための音声認識：システムの概要とリスピークの効果”, 信学技報.SP, 音声,Vol.102,No.160(20020621),pp.49-54,Jun.2002