情報処理学会論文誌 Vol. 50 No. 2 448–459 (Feb. 2009)

日本語講義音声コンテンツコーパスの作成と分析

土屋雅稔[†1] 小暮悟[†2] 西崎博光[†3]
太田健吾[†4] 山本一公[†4] 中川聖一[†4]

近年，様々な教育機関において，多数の e-Learning コンテンツが作成・利用されている．これらのコンテンツを効果的に活用して学習するには，音声データの検索および要約機能を備えた高機能なコンテンツ閲覧システムが必要である．我々が作成した日本語講義音声コンテンツコーパス（*Corpus of Japanese classroom Lecture speech Contents*）は，講義音声を対象とする認識，検索および要約などの研究を行うための基礎的な資料として設計されたコーパスであり，実際の教室で収録された音声データ，書き起こしテキスト，スライドデータからなる．本論文では，日本語講義音声コンテンツコーパスの仕様について述べると同時に，講義音声と講演音声の違い，マイク性能が講義音声認識に与える影響について分析する．

Development and Analysis of Corpus of Japanese Classroom Lecture Speech Contents

Masatoshi Tsuchiya,[†1] Satoru Kogure,[†2]
Hiromitsu Nishizaki,[†3] Kengo Ohta,[†4]
Kazumasa Yamamoto[†4] and Seiichi Nakagawa[†4]

This paper explains our developing *Corpus of Japanese classroom Lecture speech Contents*. Increasing e-Learning contents demand a sophisticated interactive browsing system for themselves, however, existing tools do not satisfy such a requirement. Many researches including large vocabulary continuous speech recognition and extraction of important sentences against lecture contents are necessary in order to realize the above system. *Corpus of Japanese classroom Lecture speech Contents* is designed as their fundamental basis, and consists of speech, transcriptions, and slides that were collected in real university classroom lectures. This paper also explains the difference about disfluency acts between classroom lectures and academic presentations.

1. はじめに

e-Learning には，学習者が自身の理解速度に合わせて学習することができ，かつ，学習する場所や時間を自由に選ぶことができるという利点があるため，様々な教育機関において e-Learning の利用が広がっている[⋆1]．e-Learning を実現するには，講義コンテンツを作成するためのシステムと，講義コンテンツを配信し，学習者の進度を記録・管理するシステムが必要である．

前者のシステムとして，たとえば，日立アドバンスドデジタル製の EZ プレゼンテータがある[⋆2]．Microsoft 社の PowerPoint 形式の講義スライドを用意し，EZ プレゼンテータを用いて講義を収録すると，講義スライドの切替え情報が記録され，講義風景の動画と講義スライドが同期した講義コンテンツを作成することができる．作成された講義コンテンツは，スライドを単位として動画再生の開始箇所を選ぶことができるようになっており，学習者は，講義中で予習・復習が必要な箇所をスライドを単位として再生し，学習することができる．しかし，学習効率をより高めるためには，学習者の求めている箇所をピンポイントに再生したり，重要箇所のみを自動的に選択再生したりする機能が必要である．また，実際の講義では，すべてのキーワードがスライドのテキストに含まれているとは限らないため，収録された音声データを検索する必要もある．このような機能を実装するには，音声ドキュメントの検索[1] や要約[2),3)]，録画内容の分析[4] といった音声ドキュメント処理技術[5] が必要である．そして，これらの技術の基盤として，講義音声を対象とする音声認識は非常に重要である．

しかし，講義音声を対象とする音声認識を実現するには，いくつかの問題がある．第 1 に，講義音声には，フィラーや言い直し，言い淀みなどの話し言葉的な現象が多数含まれている

†1 豊橋技術科学大学情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology
†2 静岡大学情報学部
Faculty of Informatics, Shizuoka University
†3 山梨大学大学院医学工学総合研究部
Department of Research Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
†4 豊橋技術科学大学情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology
⋆1 たとえば，MIT OpenCourseWare Project ⟨http://ocw.mit.edu/OcwWeb/web/home/home/index.htm⟩ など．
⋆2 http://www.hitachi-ad.co.jp/ezp/

うえに，講義内容が多岐にわたるため，発話スタイルとドメインともに適したコーパスは利用できないことがほとんどである．そのため，何らかの既存のコーパスから作成された言語モデルを，実際の講義に適応させる必要が生じる．第2に，前述のようなシステムを用いて一般の教室で収録された講義音声には，残響やノイズが大量に含まれている．また，収録には，無指向性で低音質の音声しか得られないピンマイクが用いられることが多いため，マイク性能による悪影響も大きい．つまり，様々なマイクを用いた場合の認識性能を比較し，低音質な音声を補正する技術の検討が必要である．これらの課題を研究するには，このような課題を考慮して設計・収集されたコーパスが必要不可欠である．

既存の講義音声コーパスとしては，以下のようなものがある．MIT で作成されたコーパス[6),7)]は，8 コースの 80 セミナにおける 300 時間の英語講義音声を収録したコーパスである．しかし，このコーパスは，一般的な教室環境で指向性マイクのみを用いて収録しているため，ピンマイクの性能による悪影響を評価するには不十分である．ポルトガルの LECTRA[8)–10)] プロジェクトでは，23 個のポルトガル語の講義（5.2 時間，44 k 語）を収録している．このコーパスは，ピンマイクとヘッドセットマイクの 2 つを用いて収録されており，音声認識性能と認識エラーについても分析されている．しかし，講義で用いられたスライドデータは含まれていない．

一般的な話し言葉を収録したコーパスも，講義音声認識に関する問題を部分的に解決する資源として用いることができる．2002 年に開始された Rich Transcription（RT）evaluation series [⋆1]は，最新の自動音声認識の性能を測定することができるように設計されている[11),12)]．最近の RT 評価タスクでは，音声書き起こし（Speech to Text），話者推定（Speaker Diarization），音声区間検出（Speech Activity Detection）というタスクについて，3 つのミーティングドメインに対して評価している．収録には，接話マイクが用いられている．日本語話し言葉コーパス[13)]（以下，**CSJ** と略す）は，学会講演を中心として，現代日本語の自発音声を研究用付加情報とともに大量に格納したデータベースであり，987 件の学会講演，1,715 件の模擬講演などを含む．ただし，すべての講演はヘッドセットマイクのみを用いて収録されているため，マイクによる影響は評価できない．また，講義音声と講演音声には，いくつかの点で違いがある．第 1 に，CSJ に収録されている講演の講演時間は 15 分前後と比較的短時間であるのに対して，講義は 30〜90 分とかなり長時間である．そのような長時間音声ドキュメントの要約およびコンテンツ処理を研究するには，CSJ は不十分である．第 2 に，教員は，同一の科目を複数年にわたって担当することが一般的である．そのため，前年度までの経験に基づき，時間的余裕もあるため，比較的リラックスした状態で発話することができる．それに対して，講演（特に学会講演）では，ある 1 つの内容について説明する体験は 1 度限りであり，かつ，時間的余裕は少ない．そのため，話者は，比較的緊張した状態での発話を強いられる．このような違いは，発話内容や発話スタイルなどに大きな影響を与えると予想されるが，そのような分析を行うためには，講演音声のデータベースだけでは不十分である．

本論文では，我々が作成した日本語講義音声コンテンツコーパス（*Corpus of Japanese classroom Lecture speech Contents*，以下では **CJLC** と略す）について述べる．CJLC は，講義音声を対象とする認識，検索および要約などの研究を行うための基礎的な資料として設計されたコーパスであり，実際の教室で収録された音声データ，書き起こしテキスト，スライドデータからなる．3 大学（豊橋技術科学大学，静岡大学，山梨大学）の 15 人の教員の協力を得て，数学，物理学，電気工学，情報科学などに関する 26 科目 89 講義（計 3,780 分）を収録した．

本論文の構成は以下のとおりである．2 章では，CJLC の仕様について述べる．3 章では，CJLC の収録内容を分析・検討する．4 章では，言語モデルと音声認識性能の観点から，講義音声と講演音声を比較し，マイク性能が講義音声認識に与える影響を明らかにする．最後に，5 章で結論を述べる．

2. 日本語講義音声コンテンツコーパスの仕様

先に述べたとおり，実際の講義音声を対象とする音声認識にはいくつかの問題があるが，CJLC は，その中でも特に 2 つの問題を対象として設計されている．第 1 の問題は，実際の教室における残響・ノイズ環境下において，マイク性能がどのように影響するか，という問題である．この問題を検討するため，CJLC では，複数のマイクを同時に用いて講義音声を収録している．第 2 の問題は，くだけた発話スタイルと多岐にわたる講義内容である．この問題を検討するために，CJLC では，3 大学 15 人の教員の協力を得て，数学，物理学，電気工学，情報科学などの講義を収録している．加えて，言語モデル適応[14)–17)]や音声要約[18)]，話題境界検出[19)]の研究のために必要となる付加情報として，講義に用いられたスライドデータや，教科書の表題，科目キーワードを収録している．以下では，CJLC の仕様の詳細について述べる．

⋆1 http://www.nist.gov/speech/tests/rt/

2.1 構　　成

日本語講義音声コンテンツコーパスは，複数の**講義コンテンツ**の集合である．1 つの講義コンテンツは，1 回の講義に対応し，以下の情報からなる．

- 講義情報（科目名，科目キーワード，教科書の表題，対象学年，種別[⋆1]，手段[⋆2]）
- 話者情報（年齢，性別，講義経験年数，講義担当科目数）
- 音声データ
- 書き起こしテキスト
- スライドデータ
- スライド切替えの時間情報
- 重要発話抽出情報

これらのうち，講義情報，話者情報，音声データと書き起こしテキストは必須データであり，すべての講義コンテンツに含まれる．他の 3 つは，任意データであり，一部の講義コンテンツにのみ含まれる．たとえば，黒板を用いて行う演習を収録した講義コンテンツには，スライドデータおよびスライド切替えの時間情報は含まれない．講義情報，話者情報，書き起こしテキスト，スライド切替えの時間情報および重要発話抽出情報は，XML 形式のファイルにまとめて格納している．実際のファイルを抜粋した例を**図 1** に示す．他の音声データおよびスライドデータについては，個別のファイルに格納している．

スライドデータは，実際の講義で用いられた Microsoft 製 PowerPoint 形式のファイルである．スライド切替えの時間情報は，スライドのページ番号，当該スライドの表示開始時間および表示終了時間からなる 3 つ組の列である．具体的には，図 1 中の `<Slide>` 要素がスライド 1 枚の情報に相当し，`<Slides>` 要素が講義全体の情報に相当する．時間情報は msec 単位で，日立アドバンスドデジタル製の EZ プレゼンテータを用いて記録した．

一部の講義コンテンツについては，重要発話抽出情報を付与している．情報科学の研究者 6 人に，要約率 25%の要約を得ることを目標として，重要と思われる発話を抽出するように依頼した．発話 ID が 0013 である発話が，作業者 2，作業者 3，作業者 5 の 3 人によって抽出された例を図 1 に示す[⋆3]．

2.2 録 音 条 件

実際の教材として講義音声を収録する場合には，指向性が高く高音質の音声が得られるハンドマイクやヘッドセットマイクなどではなく，無指向性で低音質の音声しか得られないピンマイクが使われることが多い．そのため，講義音声認識においては，ピンマイクによって得られた音声を補正して高精度に認識する方法は重要な課題である．このような問題意識から，1 つの講義コンテンツには，原則的に，2 種類のマイク（ピンマイクと，ハンドマイクまたはヘッドセットマイク）を用いて録音を行った 2 系統の音声データが含まれる．**表 1** に，CJLC の録音条件を示す．また，通常の教室における残響・ノイズによる影響を調査するた

```
<?xml version="1.0" encoding="UTF-8" ?>
<Lecture LectureID="L11M0030">
<Info>
  <Speaker SpeakerID="2" SpeakerBirthGeneration="35to39" SpeakerSex="男" LectureExperience="4"
    NumberOfLectures="2" />
  <Title>電子計算機応用特論 2</Title>
  <Keyword>音声認識</Keyword><Keyword>音声対話</Keyword><Keyword>マルチモーダルインタラクション</Keyword>
  <Target>修士 1 年</Target>
  <LectureType>講義</LectureType>
  <LectureMethod>スライド</LectureMethod>
</Info>
<Utterances>
  <Utterance UtteranceID="0012">
    <Time MicID="1" StartTime="101440" EndTime="103408" />
    (F ま) これから少しずつ話しますが
  </Utterance>
  <Utterance UtteranceID="0013" ExtractedBy="2,3,5">
    <Time MicID="1" StartTime="103848" EndTime="116512" />
    で (F えーと) この対話は公準に基づいてお互いに協調的に行う (F ま) 公準ていうのも次のページでまたみせます
    (F えー) 何かこう何か基準があってこれに従って普通話しているというふうに言われます
  </Utterance>
  <Utterance UtteranceID="0014">
    <Time MicID="1" StartTime="116768" EndTime="125824" />
    でその相手の感情だとかいろんなことを (F えー) 考慮して相手に発話を向けるということになります
  </Utterance>
  <Utterance UtteranceID="0015">
    <Time MicID="1" StartTime="127032" EndTime="133824" />
    で (F まあ) いろいろ (D い) 相手の意図を理解しながら話さないといけないんですけど状況がいろいろ変わりますからそ
れを考慮して
  </Utterance>
</Utterances>
<Slides>
  <Slide Page="0" StartTime="0" EndTime="22081" />
  <Slide Page="1" StartTime="22081" EndTime="86107" />
  <Slide Page="2" StartTime="86107" EndTime="170140" />
</Slides>
</Lecture>
```

図 1　講義情報，話者情報，書き起こしテキスト，スライド切替えの時間情報および重要発話抽出情報の格納形式
Fig. 1　An example of XML format.

⋆1 現時点では，「講義」「演習」の 2 通りの分類のみを行っている．
⋆2 現時点では，「スライド」「黒板」の 2 通りの分類のみを行っている．
⋆3 `UtteranceID="0013"` となっている `<Utterance>` 要素の `ExtractedBy` 属性を参照．

表 1　CJLC の録音条件
Table 1　Recording Conditions of CJLC.

マイク種別	型番	接続	録音機材	データ形式
ピンマイク	Sony ECM-T145, ECM-88B, ECM-C10	有線	PC	44.1 kHz, 192 kbps, WMA
	Panasonic WX-1800	無線		
	TOA WM-1300	無線	DAT レコーダ (Sony TCD-D8)	48 kHz, 16 bit, PCM
ハンドマイク	Sony C-355	有線		
ヘッドセットマイク	Shure SM10A	有線	IC レコーダ (Marantz PMD-671)	16 kHz, 16 bit, PCM

表 2　書き起こしテキストに用いるタグ一覧
Table 2　List of tags used in transcription texts.

タグ	概要	使用例
(F)	フィラー	(F あのー)
(D)	言い直し，言い淀みなどによる自立語の語断片	(D なん) 何デシベル
(D2)	助詞，助動詞，接辞の言い直し	(D2 が) を取り込む
(W)	漢字の読み間違えなど一時的な発音エラー	(W フインキ; 雰囲気)
(A)	アルファベットや算用数字	(A シーディーアール; CDR)
(笑)	笑いながら発話	(笑 面白い)
(泣)	泣きながら発話	(泣 どんなに)
(咳)	咳をしながら発話	(咳 えっと)
＜H＞	母音の引き延ばし	今日は ＜H＞
＜笑＞	言語音とは独立の話者の笑い声	
＜咳＞	言語音とは独立の話者の咳	
＜息＞	言語音とは独立の話者の呼吸音	
＜雑音＞	言語音とは独立の雑音	
＜フロア発音＞	話者以外のフロア (教室) からの発話	

めに，収録用の特殊な内装が施されている教室ではなく，通常の教室で収録を行っている．

2.3　書き起こしテキストの形式

すべての音声データは，音声パワーに基づいて発話区間に自動分割[20),21)]し，それぞれの発話区間に対して書き起こしを行う．

書き起こしには，音声現象を表現するために**表 2** のタグを用いる．表 2 のタグは，例外（W タグ）を除いて，CSJ の書き起こしテキストにおいて用いられているタグのサブセットであり，これらのタグの使用は，CSJ の書き起こし基準[22)]に従う．W タグは，CSJ では「転訛，発音の怠けなど一時的な発音エラー」と定義されているが，CJLC では，漢字の読み間違いなどについても W タグを用いて表現する．

書き起こしテキストを XML 形式で格納した例を図 1 に示す．1 つの発話は `<Utterance>` 要素に相当し，1 つの講義の発話全体は `<Utterances>` 要素に相当する．1 つの `<Utterance>` 要素は，発話 ID，時間情報および書き起こしテキストからなる．図 1 では，「えーと」「えー」などのフィラー部分を表現するために，F タグが用いられている．また，語断片「い」を表すために，D タグが用いられている．

3.　日本語講義音声コンテンツコーパスの分析

以下では，CJLC の収録内容を分析・検討し，話し言葉的な現象の発生頻度に注目して講義音声と講演音声の違いを明らかにする．

3.1　収録数と時間長

収録した講義音声の話者数，科目数，講義数，収録時間を**表 3** に示す．収録対象は，数学，物理学，電気工学，情報科学などの科目の講義である．それらの大分類を**表 4** に示す．表 4 より，CJLC は，情報工学系の学部および大学院において開講されている専門科目を広く収録していることが分かる．話者の年齢，講義経験年数および講義担当科目数を**表 5** に示す．表 5 より，CJLC は，年齢と講義経験年数が異なる様々な話者による発話を収録していることが分かる．このようなコーパスは，年齢や講義経験年数と話速や韻律との関係を分析するために用いることができる[23)]．発話の大部分は教員による独話である．一部の音声データには，教員と学生の質問応答形式の対話が含まれているが，書き起こしは，教員の発話部分についてのみ行っている．1 科目あたり 3.4 回の講義が収録され，1 講義あたりの平均時間長は 43.5 分である．89 講義のうち，6 講義（6.8 時間）を対象として，情報科学の研究者 6 人が重要発話抽出を行った．これは，CSJ に含まれている重要発話抽出情報（119 講演，43.5 時間）よりも少ない．しかし，CSJ において重要発話抽出の対象となっている講演の時間長は平均 13.1 分であるのに対して，CJLC において対象とした講義の平均時間長は 68 分と，比較的長時間である．このような長時間音声ドキュメントを対象とする重要発話抽出は，ほとんどなされておらず，有用なデータである．

表 6　CJLC と CSJ のデータ量
Table 6　Statistics of CJLC and CSJ.

コーパス	種別	講義数（講演数）	収録時間 [時間]	平均時間長 [秒/講義]	平均単語数 [単語数/講義]	平均タグ出現頻度 [頻度/講義]（[頻度/秒]）		
						F	D	D2
CJLC	講義	89	63.0	2,610	6,636	410.3 (0.172)	49.9 (0.0139)	3.9 (0.00194)
CSJ	学会講演	987	275.0	1,003	3,358	229.2 (0.229)	44.5 (0.0448)	3.4 (0.00343)
	模擬講演	1,715	330.6	694	2,122	118.8 (0.169)	26.0 (0.0370)	1.4 (0.00201)
	対話	58	12.3	765	2,613	322.2 (0.420)	43.9 (0.0588)	1.4 (0.00195)

表 3　話者数・科目数・講義数・収録時間
Table 3　Number of speakers/courses/lectures and duration.

話者数	15　（全員が男性）
科目数	26　（数学，物理学，電気工学，情報科学など）
講義数	89　（うち，スライドデータを含む講義は 62 講義）
収録時間	3,780 分

表 4　科目・講義の分類
Table 4　Categories of courses and lectures.

大分類	科目数	講義数	例
理学系	7	10	線形代数学，物理学実験
電気工学系	2	6	基礎電気理論
情報工学系	17	73	プログラミング言語論

表 5　話者の年齢・講義経験年数・講義担当科目数
Table 5　Ages of speakers, their teaching history and number of their courses.

	最小	平均	最大
収録時の年齢	31	45.7	58
収録時の講義経験年数	2	14.2	30
収録年度の講義担当科目数	2	4.2	7

講義時間長に基づいて講義を集計した結果を**図 2** に示す．先に述べたとおり，全講義を平均すると講義時間長は 43.5 分であるが，図 2 より明らかに，講義時間長の観点からは 2 種類の講義がある．以下の議論で必要がある場合には，60 分未満の講義を**短講義**，60 分以上の講義を**長講義**と呼んで区別する．短講義の大部分は，実験や演習の手順説明やガイダンスなどである．それに対して，長講義は，通常の教室で行われる授業形式の講義などである．このように，CJLC は，長講義および短講義の双方を収録しており，時間長がフィラーや言い直し，言い淀みに与える影響を分析することが可能である（3.2 節）．

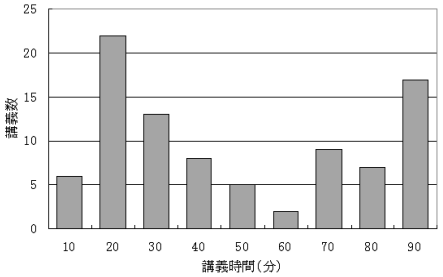
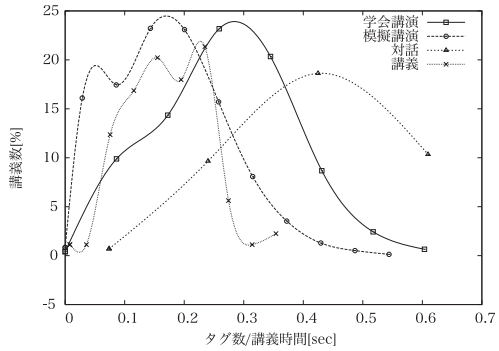


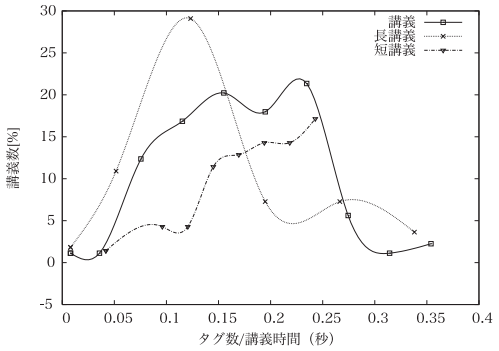
図 2　講義時間と講義数
Fig. 2　Number of lectures and durations.

次に，1 回の講義（または講演）あたりの時間長の観点から，講義音声と講演音声を比較する．講演音声データとしては，日本語話し言葉についての最大のコーパスである CSJ を用いる．CSJ には，(1) 学会講演，(2) 模擬講演，(3) 対話，(4) 朗読という 4 種類の音声データが含まれているが，本論文では最初の 3 つのみを用いる．学会講演は，理工学，社会，人文などの領域の 9 つの学会における講演を収録したデータである．模擬講演は，一般話者による日常的話題[⋆1]についての 10 分から 15 分程度のスピーチである．対話は，学会講演インタビュー，模擬講演インタビュー，課題指向対話および自由対話からなる．講義音声と講演音声の比較を**表 6** に示す．表 6 より，平均時間長および平均単語数で比較すると，CJLC の講義音声は，CSJ の講演音声よりも長時間の独話音声であることが分かる．講義音声を短講義と長講義に分類した場合でも，ほとんどの講義は，講演音声の平均時間長よ

⋆1 CSJ の作成者によって，あらかじめ 3 種の一般的テーマ（たとえば，「私の人生でもっとも嬉しかったこと」）が指定されている．

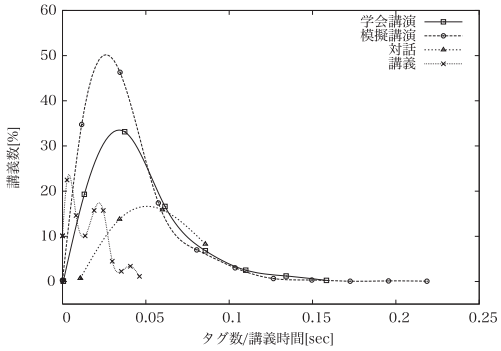


(a) 講義音声と講演音声の比較

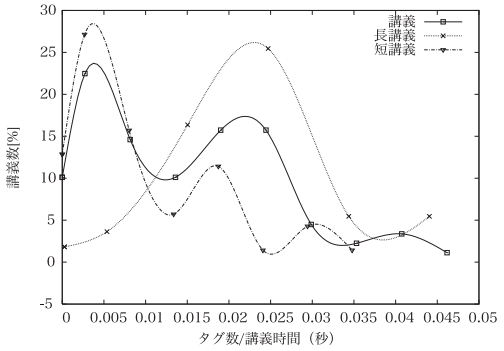


(b) 短講義と長講義の比較

図 3　タグ F の出現頻度
Fig. 3　Distribution of Tag F.



(a) 講義音声と講演音声の比較



(b) 短講義と長講義の比較

図 4　タグ D の出現頻度
Fig. 4　Distribution of Tag D.

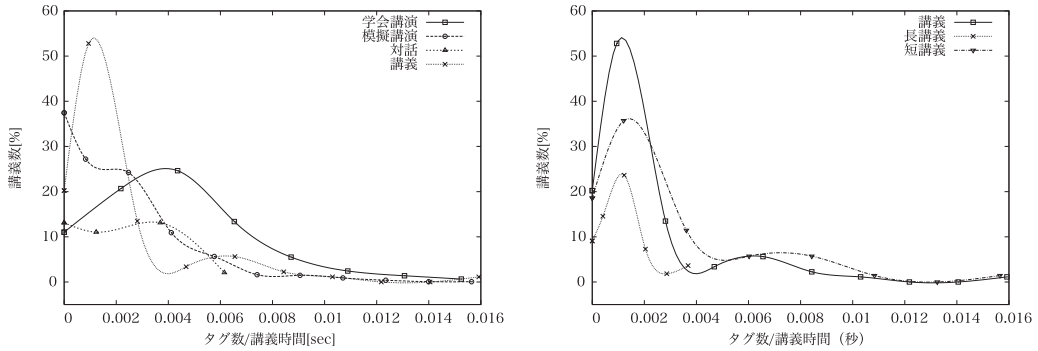
りも長い．この時間長の違いが，フィラーや言い直し，言い淀みに与える影響については，次節で検討する．

3.2　フィラー・言い直し・言い淀みの発生頻度

本節では，フィラー，言い直し，言い淀みの発生頻度の観点から，講義音声と講演音声を分析する．

まず，講義音声，学会講演，模擬講演および対話におけるフィラー（タグ F）の時間あたり出現頻度[⋆1]を**図 3** (a) に示す．図 3 (a) より，講義音声は，フィラーの出現頻度の観点から

⋆1 異なるサンプル数の出現分布を相互に比較できるように，近似的に，頻度数 [%] を 10 区間相当数に正規化している．



(a) 講義音声と講演音声の比較　　(b) 短講義と長講義の比較

図 5　タグ D2 の出現頻度
Fig. 5　Distribution of Tag D2.

は，講演音声（学会講演および模擬講演）に近いことが分かる．短講義と長講義のフィラーの出現頻度の比較を図 3 (b) に示す．短講義は，長講義に比べて，フィラーの出現頻度が僅かに大きいようである．しかし，短講義，長講義のフィラーの出現頻度分布は，講演音声のフィラーの出現頻度分布の範囲内に収まっており，大きな差ではないと思われる．

講義音声，学会講演，模擬講演および対話における言い直しまたは言い淀み（タグ D およびタグ D2）の時間あたり出現頻度を**図 4** (a) および **図 5** (a) に示す．図 4 (a) および図 5 (a) より，講義音声は，講演音声（学会講演および模擬講演）よりも，言い直しまたは言い淀みの出現頻度が小さい．その原因はおそらく，講義音声と講演音声の時間長および発話状況の違いにあると思われる．つまり，講演音声（特に学会講演）は，時間的に厳しく制限され緊張した状況での発話であるために，講義音声よりも，言い直しまたは言い淀みの発生頻度が高くなると考えられる．図 4 (b) および図 5 (b) より，自立語の言い直しまたは言い淀み（タグ D）の出現頻度は短講義と長講義で異なるが，機能語の言い直し（タグ D2）の出現頻度は短講義と長講義でほとんど異ならない．

以上より，講義音声と講演音声の違いは，フィラーの出現頻度ではなく，言い直しまたは言い淀みの出現頻度に現れている．

4. 言語モデルと音声認識による評価

以下では，CJLC の収録内容を，言語モデルと音声認識性能の観点から，講義音声と講演音声を比較し，加えて，マイク性能が講義音声認識に与える影響について述べる．

4.1 言語モデルによる比較

本節では，発話スタイルとドメインの点から，CJLC に収録されている講義音声を分析する．全 89 講義のうち，**表 7** に示す 11 講義を特に分析対象とする[⋆1]．なお，表 7 の最下行（CSJ）は，CSJ における男性話者 10 講演の音声認識用テストセットから 4 講演（A01M0007, A01M0035，A01M0074，A05M0031）を選択したものであり，比較対象として用いる．

比較検討のために作成した言語モデルの仕様を，**表 8** に示す．$\mathrm{CSJ}_{970-20k}$ モデルは，CSJ の学会講演のみを学習コーパスとして作成した言語モデルである．この学習コーパスには，音響学会などで行われた音声言語処理に関する学会講演が含まれているため，その領域の講義とドメインが一致していることが期待される．$\mathrm{CSJ}_{3300-20k}$ モデルおよび $\mathrm{CSJ}_{3300-40k}$ モデルは，CSJ に収録されている学会講演，模擬講演および対話を学習コーパスとして作成した言語モデルである．これらは，話し言葉的な現象を広くカバーしていることが期待される．最後に，NEWS_{20k} モデルおよび NEWS_{40k} モデルは，毎日新聞（1991 年〜2004 年）から作成した言語モデルである．

⋆1 この 11 講義は，2008 年 1 月時点で音声データの書き起こしが完了していた講義である．

表 7　比較対象として選択した講義一覧
Table 7　List of lectures for experiment.

講義 ID	話者 ID	文数	時間 [秒]	フィラー率 [%]	科目名	キーワード
L1	S1	269	1,213	6.80 (263/3,869)	電子計算機応用特論 2	音声言語処理
L2		236	1,274	4.54 (179/3,946)		DP マッチング
L4	S2	558	937	14.75 (496/3,364)	電子計算機応用特論 1	自然言語処理
L6	S3	1,480	3,623	6.64 (1,006/15,157)	ソフトウェア工学	プログラムデザイン，コーディング
L7	S4	743	1,903	8.20 (484/5,901)	物理学実験	ダイオード，P 型半導体，N 型半導体
L8	S5	1,163	4,193	5.69 (672/11,803)	アルゴリズムとデータ構造 1 および演習	二分木
L9		903	3,115	6.57 (550/8,367)		バブルソート
L10		820	3,285	8.24 (745/9,037)		選択ソート，挿入ソート
L11		564	2,261	7.72 (504/6,529)		クイックソート
L12	S1	212	160	4.70 (80/1,702)	電子計算機応用特論 2	言語モデル
L13	S2	311	254	10.09 (187/1,853)	電子計算機応用特論 1	自然言語処理
CSJ		1,771	4,568	10.91 (2,050/18,793)	（音響学会講演など）	

表 8　言語モデル
Table 8　Specifictions of language models.

	語彙サイズ	学習コーパス	
		講演数（記事数）	サイズ [Byte]
$CSJ_{970-20k}$	20 k	970 ⋆1	23 M
$CSJ_{3300-20k}$	20 k	3,285 ⋆2	123 M
$CSJ_{3300-40k}$	40 k		
$NEWS_{20k}$	20 k	1,499,936	1,400 M
$NEWS_{40k}$	40 k		

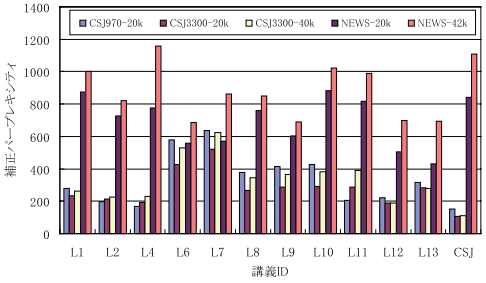


図 6　補正パープレキシティ
Fig. 6　Adjusted Perplexity.

図 6 に講義音声の補正パープレキシティ⋆3を，**図 7** に未知語率を示す．L1，L2，L12，L13 は，CSJ から作成された言語モデル（$CSJ_{970-20k}$ モデルなど）の補正パープレキシティが小さく，さらに，未知語率も低いことから，発話スタイル，ドメインともに CSJ に似ていると考えられる．L6，L7，L8，L9，L10，L11 は，いずれも非常に未知語の多い講義であり，ドメインの点で CSJ とは異なる講義群である．CSJ から作成された言語モデルの補正パープレキシティの点から見ると，これらの講義はさらに，L6，L7 からなる講義群と，L8〜L11 からなる講義群の 2 つに分類できる．これらは，発話スタイルの点で CSJ に似ている講義群（L8〜L11）と，似ていない講義群（L6，L7）と考えられる．L4 は，例外的に，未知語中のフィラー率が非常に高い講義である（**図 8**）．

新聞から作成した $NEWS_{20k}$ モデルおよび $NEWS_{40k}$ モデルと CSJ から作成したモデルを比較すると，すべての講義について，補正パープレキシティおよび未知語率の両方の観点で，新聞から作成したモデルは不十分であることが分かる．これは，発話スタイルおよびド

⋆1 学会講演のみからなる．
⋆2 CSJ に含まれるすべての講演からなる．
⋆3 補正パープレキシティは，テストコーパス中に出現した未知語率を考慮した尺度である．テストコーパス中に出現した未知語の延べ頻度を o，異なり数を m，総単語数を n とすると，補正パープレキシティ PP^* は，パープレキシティ PP を用いて次式で定義される[24]．

$$\log_2 PP^* = \log_2 PP + \frac{o}{n} \log_2 m$$

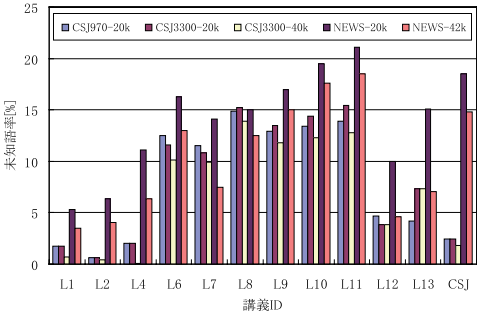


図 7　未知語率
Fig. 7　OOV ratio.

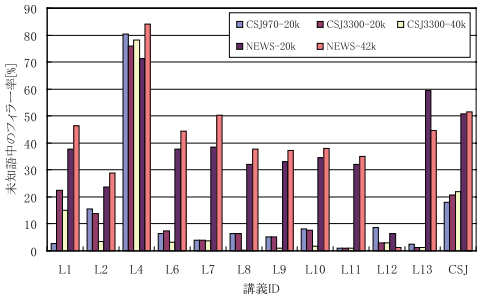


図 8　未知語中のフィラー率
Fig. 8　Filler ratio in OOV.

メインの双方について，言語モデルの適応が必要であることを示唆している．

各講義および講演音声（CSJ）の書き起こしと，新聞記事（毎日新聞 1993 年）を対象として，英文字，数字，カタカナのみからなる語の出現率を調べた結果を**図 9** に示す．L1〜L13 を平均すると，英文字，数字，カタカナのみからなる語の出現率は，それぞれ 1.6%，2.3%，3.7%である．図 9 より，ソフトウェアに関する講義（L6）と演習（L8〜L11），および言語モデルに関する講義（L12）では，講演音声および新聞記事よりも，カタカナ語の出現率が高くなっている．しかも，これらの 6 講義に出現したカタカナ語の 28%は未知語である．DP マッチングに関する講義（L2），物理学実験（L7），およびソフトウェアに関する演習（L8〜L11）では，英文字のみからなる語の出現率が高くなっている．DP マッチングに関する講義（L2）では，数字のみからなる語の出現率も高くなっている．このように，講義音声においては，講義の主題によって，これら 3 種の語の出現率が大きく変化する．また，講義音声全体を見ても，講演音声および新聞記事よりも，カタカナ語および英文字のみからなる語の出現率が高くなっており，音声認識を困難にする要因となっていると考えられる．

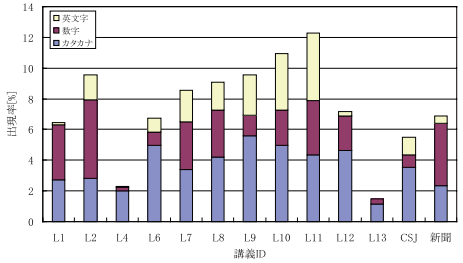


図 9　英文字，数字，カタカナのみからなる単語の出現率
Fig. 9　Ratio of words consisting of alphabets/numericals/katakanas.

表 9　音響分析条件
Table 9　Conditions of acoustic analysis.

量子化ビット数	16 bit
サンプリング周波数	16 KHz ⋆1
分析窓	Hamming 窓
分析窓長	25 ms
分析窓間隔	10 ms
特徴パラメータ	MFCC (12 次元) + ΔMFCC (12 次元)
	+ΔΔMFCC (12 次元)
	+ΔPOW (1 次元) + ΔΔPOW (1 次元)

4.2　音声認識性能の比較

本節では，音声認識性能の観点から，講義音声と講演音声を比較する．

認識に用いる音響モデルは，CSJ 学会講演音声のうち，男性によって行われた 797 講演から学習した 38 次元トライフォンモデルである．混合数は 32 混合で，状態数は 3,013 状態である．音声認識器には Julius rev. 3.5.3 を用いた．音響分析条件は**表 9** のとおりであ

⋆1 48 kHz で収録した音声は 16 kHz にダウンサンプリングした．

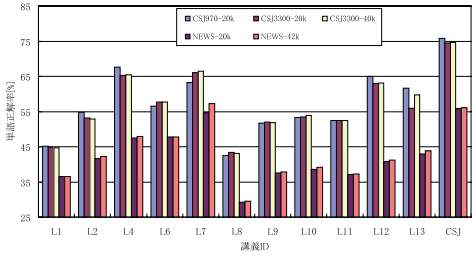


図 10　単語正解率[⋆1]
Fig. 10　Word correct ratio.

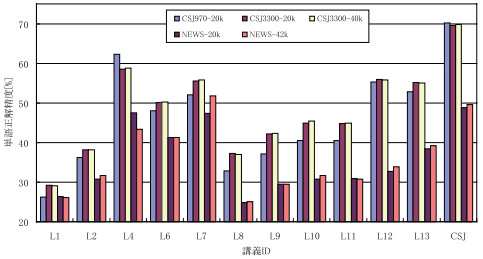


図 11　単語正解精度[⋆1]
Fig. 11　Word accuracy ratio.

る．言語モデルには，表 8 と同じ 5 種類のモデルを用いた．講義音声の収録にあたっては，無指向性で低音質の音声しか得られないピンマイクが用いられることが多いため，マイク性能による悪影響が大きい．現実的な状況における音声認識性能を比較するため，ピンマイク（表 1）を用いて収録した講義音声データを用いた．

図 10 に単語正解率を，**図 11** に単語正解精度を示す．CSJ から作成された言語モデル（$\mathrm{CSJ}_{970-20k}$ モデルなど）を用いて音声認識を行った場合，講義音声では，講演音声よりも 8%以上性能が悪化している．実験対象の 11 講義のうち，CSJ から作成された言語モデルでの未知語率が CSJ よりも低い講義は，L1，L2，L4 の 3 つである（図 7）．さらに，この 3 講義は，CSJ から作成された言語モデルによって求められた 3 通りの補正パープレキシティが小さい講義群に属していることから，相対的に認識の容易な講義と考えられる．しかし，これらの講義も，CSJ から作成された言語モデルでの補正パープレキシティはかなり大きく，他の講義と同様に，低い音声認識結果しか得られていない．新聞から作成された言語モデル（NEWS_{20k} モデルなど）を用いて音声認識を行った場合も，例外（L7）を除いて，講演音声よりも 8%以上性能が悪化している．このように，講義音声の認識は，講演音声の認識よりも相当に難しい．その理由としては，マイク性能による悪影響，言語モデルにおけるドメインの不一致，発声スタイルの違いなどが考えられる．マイク性能による悪影響については，次節で検討する．言語モデルについては，スライドデータや科目キーワードを用いた言語モデルのドメイン適応が有効である[14),16)]．

⋆1 L1〜L13 はピンマイク（表 1）で収録した講義音声データ，CSJ はヘッドセットマイク（CROWN CM–311a）で収録された講演音声データを，それぞれ用いた実験結果である．

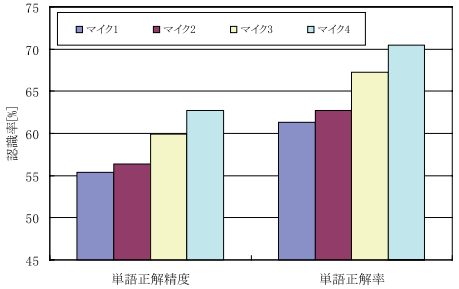


図 12　マイク性能の比較
Fig. 12　Compare of mic performances.

4.3　マイク性能が講義音声認識に及ぼす影響

本節では，講義音声認識において，講義を収録するマイクが認識率に影響を与えることを示す．CJLC は，通常は 2 系統の音声データを含むが，L12，L13 については，特別に 4 つのマイクを用いて収録を行った．使用したマイクは以下の 4 種類である．

- マイク 1：ECM-C10（SONY）標準的なピンマイク
- マイク 2：ECM-88B（SONY）高性能ピンマイク
- マイク 3：ECM-355（SONY）ハンドマイク
- マイク 4：ISOMAX ヘッドセットマイク

言語モデルとして $\mathrm{CSJ}_{3300-40k}$ を用いた場合の実験結果を**図 12** に示す．なお，図 12 の単語正解率と単語正解精度は，L12 および L13 全体に対する値である．マイク性能の違いを見ると，標準的ピンマイクと高性能ピンマイクにおいて 1%程度の差があり，ピンマイクとハンドマイクにおいては 3〜4%の差があり，また，ハンドマイクに比べてヘッドセット

マイクの方が約 3%認識率が良いことが分かった．この認識率の差は，ハンドマイクでは，口とマイクの間の距離を一定に保つことが難しいが，ヘッドセットマイクは，口とマイクの間の距離がつねに一定に保たれ，質の良い音声が収集できることに起因していると考えられる．しかし，CSJ の講演を対象とした場合の認識性能と比較すると，まだかなり悪い．

5. おわりに

本論文では，我々が作成した日本語講義音声コンテンツコーパス（CJLC）について述べた．講義音声を対象とする認識，検索および要約などの研究を行うための基礎的な資料として設計されたコーパスであり，実際の教室で収録された音声データ，書き起こしテキスト，スライドデータからなる．音声データは複数のマイクを同時に用いて収録されており，マイク性能の影響を評価することができるようになっている．本論文では，CJLC の収録内容を分析・検討し，講義音声と講演音声の違い，マイク性能が講義音声認識に与える影響を明らかにした．

今後の課題として，発話状況の違いが発話スタイルや話速，韻律に与える影響，余談や冗長部分の有無などのような講義音声の特徴をさらに分析することを予定している．また，ピンマイクで収録された低音質の音声に何らかの補正を加えて高精度な音声認識を実現する方法を検討する．CJLC のモニタ版は，http://www.slp.ics.tut.ac.jp/CJLC/で公開している．収録したデータを整理し，できるだけ早期に CJLC の正式版を公開する予定である．

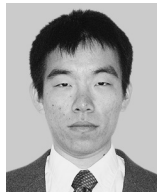
謝辞　講義の収録に協力してくださった教員各位に感謝します．本研究の一部は，総務省戦略的情報通信研究開発推進制度「音声ドキュメントのセマンティックコンテンツ化と音声対話による高度利用化の研究」によるものです．

参考文献

1) Nishizaki, H. and Nakagawa, S.: Japanese Spoken Document Retrieval Considering OOV Keywords Using LVCSR System with OOV Detection Processing, *Proc. HLT2002*, pp.144–151 (2002).
2) Hori, C., Hori, T. and Furui, S.: Evaluation method for automatic speech summarization, *Proc. EUROSPEECH2003*, pp.2825–2828 (2003).
3) Togashi, S., Yamaguchi, M. and Nakagawa, S.: Summarization of spoken lectures based on linguistic surface and prosodic information, *Proc. IEEE/ACM Workshop on Spoken Language Technology* (*SLT*), pp.34–37 (2006).
4) Li, Y. and Dorai, C.: Instructional Video Content Analysis Using Audio Information, *IEEE Trans. Audio, Speech and Language Process.*, Vol.14, No.6, pp.2264–2274 (2006).
5) 中川聖一：音声ディクテーションから音声ドキュメント処理へ，日本音響学会 2007 年秋季研究発表会講演論文集，pp.1–4 (2007).
6) Park, A., Hazen, T.J. and Glass, J.: Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling, *Proc. ICASSP2005*, pp.497–500 (2005).
7) Glass, J., Hazen, T.J., Hetherington, L. and Wang, C.: Analysis and Processing of Lecture Audio Data: Preliminary Investigations, *Proc. Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, pp.9–12 (2004).
8) Lamel, L., Adda, G., Bilinski, E. and Gauvain, J.: Transcribing Lectures and Seminars, *Proc. EUROSPEECH2005*, pp.1657–1660 (2005).
9) Trancoso, I., Nunes, R., Neves, L., Vianan, C., Moniz, H., Caseiro, D. and Mata, A.I.: Recognition of Classroom Lectures in European Portuguese, *Proc. Interspeech2006-ICSLP*, pp.281–284 (2006).
10) Trancoso, I., Martins, R., Moniz, H., Mata, A.I. and Viana, M.: The LECTRA Corpus—Classroom Lecture Transcriptions in European Portuguese, *Proc. LREC2008* (2008).
11) Fügen, C., Wölfel, M., McDonough, J.W., Ikbal, S., Kraft, F., Laskowski, K., Ostendorf, M., Stüker, S. and Kumatani, K.: Advances in Lecture Recognition: The ISL RT-06S Evaluation System, *Proc. Interspeech2006-ICSLP*, pp.1229–1232 (2006).
12) Fügen, C., Kolss, M., Bernreuther, D., Paulik, M., Stüker, S., Vogel, S. and Waibel, A.: Open Domain Speech Recognition & Translation: Lectures and Speeches, *Proc. ICASSP2006*, pp.569–572 (2006).
13) Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *Proc. SSPR2003*, pp.7–12 (2003).
14) 富樫慎吾，北岡教英，中川聖一：スライド情報を用いた言語モデル適応による講義音声認識，日本音響学会 2006 年春季研究発表会講演論文集，pp.191–192 (2006).
15) 根本雄介，秋田祐哉，河原達也：講義音声認識のためのスライド情報を用いた言語モデル適応，第 1 回音声ドキュメント処理ワークショップ講演論文集，pp.89–94 (2007).
16) 小暮　悟，西崎博光，土屋雅稔，富樫慎吾，山本一公，中川聖一：日本語講義音声コンテンツコーパスの構築と講義音声認識手法の検討，第 2 回音声ドキュメント処理ワークショップ講演論文集，pp.7–14 (2008).
17) 徳田　翔，西崎博光，関口芳廣：講義音声認識のための WEB 文書を用いた言語モデルの適応化と語彙選択，第 2 回音声ドキュメント処理ワークショップ講演論文集，pp.97–104 (2008).
18) 富樫慎吾，山口　優，北岡教英，中川聖一：講義音声の認識・要約・インデックス化

の検討，情報処理学会研究報告，Vol.2006–SLP–73, pp.57–62 (2006).
19) 大久保崇，菊池英明，白井克彦：音声対話における韻律を用いた話題境界検出，情報処理学会研究報告，Vol.2003–SLP–124, pp.235–240 (2003).
20) 北岡教英，山田武志，柘植　覚，宮島千代美，西浦敬信，中山雅人，傳田遊亀，藤本雅清，山本一公，滝口哲也，黒岩眞吾，武田一哉，中村　哲：CENSREC-1-C：雑音下音声区間検出評価基盤の構築，情報処理学会研究報告，Vol.2006–SLP–107, pp.1–6 (2006).
21) Otsu, N.: A threshold selection method from gray-level histograms, *IEEE Trans. Systems, Man and Cybernetics*, Vol.SMC-9, No.1, pp.62–66 (1979).
22) 小磯花絵，間淵洋子，西川賢哉，斎藤美紀，前川喜久雄：転記テキストの仕様 Version 1.0，CSJ 付属文書 (2003).
23) 小林健司，宗宮充宏，名取　賢，西崎博光，関口芳廣：講義音声の自動評価のための各種特徴量の調査，第 2 回音声ドキュメント処理ワークショップ講演論文集，pp.143–148 (2008).
24) 中川聖一，赤松裕隆：未知語を含む文集合のパープレキシティの算出法—新補正パープレキシティ，日本音響学会研究発表会講演論文集，pp.63–64 (1998).

(平成 20 年 6 月 4 日受付)
(平成 20 年 11 月 5 日採録)



土屋　雅稔（正会員）

1998 年京都大学工学部卒業．2004 年京都大学大学院情報学研究科知能情報学専攻博士課程単位認定退学．博士（情報学）．2004 年豊橋技術科学大学情報処理センター助手．2007 年より同大学情報メディア基盤センター助教．自然言語処理に関する研究に従事．言語処理学会会員．

小暮　　悟

2002 年豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻修了．同年豊橋技術科学大学工学部研究員．同年愛知教育大学教育学部情報教育講座助手．2004 年静岡大学情報学部情報科学科助手．現在，同大学同学科助教．音声対話システム，音声インタフェース，プログラミング学習者・教師支援に興味を持つ．博士（工学）．日本音響学会，電子情報通信学会，人工知能学会各会員．

西崎　博光

1998 年豊橋技術科学大学工学部卒業．2003 年豊橋技術科学大学院工学研究科博士後期課程修了．博士（工学）．2003 年山梨大学大学院医学工学総合研究部助手．現在，山梨大学大学院医学工学総合研究部助教．音声言語処理，特に音声ドキュメント検索，音声の自動評価に関する研究に従事．電子情報通信学会，日本音響学会各会員．

太田　健吾（学生会員）

2007 年豊橋技術科学大学工学部卒業．現在，豊橋技術科学大学院工学研究科情報工学専攻在学．音声言語処理に関する研究に従事．日本音響学会，電子情報通信学会，人工知能学会各学生会員．

山本　一公（正会員）

1995 年豊橋技術科学大学工学部卒業．2000 年豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻修了．博士（工学）．2000 年信州大学工学部助手．2007 年より豊橋技術科学大学情報工学系助教．音声情報処理（主に音声認識）に関する研究に従事．日本音響学会，電子情報通信学会各会員．

中川　聖一（フェロー）

1976 年京都大学大学院博士課程修了．同年京都大学情報工学科助手．1980 年豊橋技術科学大学情報工学系講師．1990 年教授．1985〜1986 年カーネギメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．工学博士．1977 年電子通信学会論文賞，1988 年 IETE 最優秀論文賞，2001 年電子情報通信学会論文賞，各受賞．電子情報通信学会フェロー．情報処理学会フェロー．著書『確率モデルによる音声認識』（電子情報通信学会編），『音声聴覚と神経回路網モデル』（共著，オーム社），『パターン情報処理』（丸善），『Spoken Language Systems』（編著，IOS Press）等．