Vol.2011-CLE-4 No.4
2011/5/13

講義ビデオの活用に向けた講義音声の発話特徴分析

永井孝幸[†1]　豊田寿行[†2]　名古屋孝幸[†2]
西澤弘毅[†2]　今井正和[†2]

著者らは市販ハイビジョンカメラを用いた講義自動収録システムを開発し，毎週 20 科目以上の講義を日常的に収録している．収録したビデオの活用方法の一つとして，教員自身がビデオを見直して講義の改善に役立てることが考えられる．しかし，ビデオ全体を見直すことは時間的に困難であることから，講義の状況を手早く把握するための工夫が必要になる．本報告では，講義音声の発話特徴から講義状況を把握する試みについて述べる．

Analisys of speech characteristics in daily-recorded lecture videos and their application to finding distinctive lectures

Takayuki Nagai,[†1] Toshiyuki Toyota,[†2]
Takayuki Nagoya,[†2] Koki Nishizawa[†2]
and Masakazu Imai [†2]

To make lecture capture a daily educational tool in higher educational organization, we have developed an automated lecture capture system using off-the-shelf high-definition camcorder and have been recording more than 20 lectures every week. One possible application of these videos is to use them as a tool for improving lectures. Reviewing captured videos by lecturers themselves is a good way of faculty development. However, watching entire videos is time-consuming and difficult to be accepted by most of lecturers. We need an effective tool to summarize how lecture was given. In this report, we pay attention to speech characterisics of lecturers and use them to classify lectures and find outliers.

1. はじめに

近年，OpenCourseWare や iTunes U を初めとして大学の講義を収録・配信する取り組みが盛んであり，Academic Earth[⋆1] のようにインターネット上に存在する質の高い講義ビデオを整理してオンライン教育の場として提供する動きも始まっている．講義ビデオはこのような学外に対する知の配信の手段としてだけでなく，予復習を初めとする学生の教育支援や教育改善のツールとしても有用であり，活用の範囲は広い．

そこで著者らは市販ハイビジョンカメラを用いた講義自動収録システムを開発し，毎週 20 科目以上の講義を日常的に収録してきた[1)–3)]．収録したビデオの活用方法の一つとして，教員自身がビデオを見直して講義の改善に役立てることが考えられる．しかし，継続的に大量に収録したビデオ全体を教員自身が見直すことは時間的に困難であることから，短時間で講義の状況を把握できるようにするための仕組みが不可欠である．

講義音声を機械的に分析する試みとしては，連続音声認識の立場からの研究が多数行われている．土屋らの研究では実際の講義で収録された音声に対して，音声認識のための言語モデルの基礎となる日本語講義音声コンテンツコーパスの作成[4)] を行っている．また，宗宮[5)] らは音声に含まれる「フィラー (間投詞)」と音声認識率との関連について研究を行っている．西村らの研究[6)] では放送大学の講義音声を対象として自由発話コーパスを作成し，音声認識時の単語誤り率を改善している．西崎らは教員の音声から話速・声量・明瞭性・抑揚・ポーズの 5 つの特徴を抽出し，講義音声評価アンケートとの関連を調査している[7)]．

講義ビデオアーカイブの構築・活用の立場からも音声分析の研究が行われており，例えば山口らの研究[8)] では音声認識によりキーワードを抽出し，動画検索用データベースの構築を試みている．自然発話コーパスの整備が進んでいる英語についてはすでに商用での音声認識サービスが始まっており，例えば商用の講義収録システム Echo360 ではクラウド型の音声認識サービスの提供を行っている[⋆2]．また，米 Yap 社も商用のクラウド型音声認識サービスを提供している[⋆3]．

†1 熊本大学総合情報基盤センター
Center for Multimedia and Information Technologies, Kumamto University
†2 鳥取環境大学環境情報学部
Department of Environmental Studies, Tottori University of Environmental Studies
⋆1 http://academicearth.org/
⋆2 http://echo360.com/news-events/press-releases/pr073009/
⋆3 http://yapme.com/speech-cloud.html

Vol.2011-CLE-4 No.4
2011/5/13

表 1 分析対象とした講義音声
Table 1 Summary of the analyzed sounds

収録地点	収録方法	対象	収録回数	収録時間(分)
鳥取	自動	対面講義(17科目)	193	18,240
熊本	自動	対面講義(2科目)	59	5,285
	手動	対面講義(2科目)	6	345
合計			258	23,870

本報告では，話者が固定されており安定した分析が可能な講師マイク音声を対象に，発話特徴の分析と特徴的な講義の検出を試みる．今回分析の対象とする講義音声は実際の講義室で自然発話された大量の未編集音声であり，利用した音響設備が講義によって異なるだけでなく，講師発話以外の雑音 (物音，学生の声，音響ノイズなど) も含まれている．

今回の分析では発話区間の出現状況と音声認識結果に基づいて講義全体の傾向を抽出し，全体的な傾向から外れた特徴を持つ音声を検出することで講義の特徴付けを行った．講義の時間配分や進捗状況，学科全体での講義実施状況を把握するには発話内容そのものでは詳細すぎるため，発話区間の出現状況にもとづいた分析を中心にしている．

本報告の構成は以下の通りである．まず 2 節で今回分析の対象とする講義音声データの概要について述べる．次に，分析にあたって音声データに施した前処理について 3 節で述べる．発話区間の出現状況に基づく分析結果を 4 節で，音声認識結果に基づく分析結果を 5 節で述べる．最後の 6 節では，今回の分析結果に基づき，講義音声の発話特徴を分析することで，講義ビデオのどのような応用が可能になるか考察する．

2. 分析対象講義音声データの概要

分析の対象とするのは鳥取環境大学および熊本大学での対面講義 (1 コマ 90 分) を市販のハイビジョンカメラを用いて自動収録システム[3] および手動撮影により収録したものである．収録期間は 2010 年度後期 (2010 年 9 月末から 2011 年 1 月下旬) であり，対象科目数は 21 科目，講義数は 258(外部講師により実施された講義は除く)，対象となる講師は 18 名である (表 1)．

自動収録システムによる音声収録では，講義室の有線/無線ハンドマイクの音声をミキサー経由でカメラのマイク入力に接続したものを録音している．手動撮影による収録では，Bluetooth ワイヤレスマイクを講師に装着してもらい音声の収録を行っている．いずれの収録においても，収録時の音声は 48000Hz,16bit ステレオの AC3 形式で保存を行った．

収録音声には講師音声以外の雑音も含まれており，無発話時に機器が生成する定常ノイズをはじめとして，板書時のペンの音や教科書をめくる音，衣擦れの音などの講師が発する発話以外の音や，講義室前方に着席している学生が発する音などの雑音も含まれている．

講義を収録することは講師に事前に伝えてあり，講師によっては講義中一時的に講師マイクを OFF にし，オフレコ発言をするケースもある．また，講師がマイク・音響の電源を入れることを忘れたり，あるいは，講師マイクなしで講義を続けるなどして講義音声がうまく収録されていないケースも含まれている．

3. 講義音声分析の前処理

講義における発話状況を分析するため，前処理として，これまで著者らが用いてきた手法[9] を用いて各講義音声に対して 0.1 秒単位で発話区間/無音区間の判定を行った．平均以上の強さを持った音声振幅が一定時間継続するかどうかで判定を行っており，紙をめくる音や板書時のペンの音など，持続時間の短い雑音についてはこの段階で取り除かれる．一方，衣擦れの音など比較的音量が大きくまた持続時間の長い雑音については発話区間として誤検出される傾向がある．しかし今回は解析の対象とする音声データの量が多いため，発話の全体的な傾向を見る際にはこれら誤検出の影響は小さいと考えられる．

発話解析において発話の区切りとする無音区間の長さを決めるため，講義音声に出現する無音区間の長さの分布を求めたのが図 1 である．今回の分析では正確な発話数を求める必要はなく講義間の相対的な比較ができればよい．そこで，どの講師の音声にも共通して大量に含まれる極めて短い無音区間を除外できればよいものとし，発話区切りとして用いる無音区間の長さの閾値には 0.6 秒を用いた．これは音声に含まれる無音区間のうち約 70%の区間を無視することに対応する．本報告の分析では，特に断りのない場合，0.6 秒以上の無発話区間が出現した位置で発話区間を区切るものとする (図 2)．

以降，(前もって定めた長さ以上の) 無発話区間で分割済みの発話区間のうち，無音区間を 1 つも含まないものを**連続発話区間**，無音区間を 1 つ以上含むものを**不連続発話区間**と呼んで区別する．通常，ひと続きの発話の中には短い無音区間が多数含まれるため，自然な発話には不連続発話区間が対応する．実際の講義音声では，1 秒以上の長さを持つ連続発話区間は「えーーー」などの長く引き延ばす発声や，切れ目のない背景雑音 (音響ノイズ，学生の雑談など) に対応している．

4. 発話区間にもとづく講義状況の集計

講義音声中の発話区間を分析し，出現位置・頻度を集計することで，講義音声を聞き直す

区間の長さ(秒)	頻度	累積 %
0.1	216760	31.00%
0.2	120706	48.27%
0.3	56369	56.33%
0.4	38008	61.77%
0.5	36105	66.93%
0.6	28861	71.06%
0.7	23740	74.45%
0.8	19854	77.29%
0.9	16850	79.70%
1.0	14318	81.75%
1.1	12054	83.47%
1.2	10582	84.99%
1.3	8952	86.27%
1.4	7696	87.37%
1.5	6838	88.35%
1.6	5965	89.20%
1.7	5219	89.95%
1.8	4608	90.61%
1.9	4214	91.21%
2.0	3680	91.73%
次の級	57790	100.00%

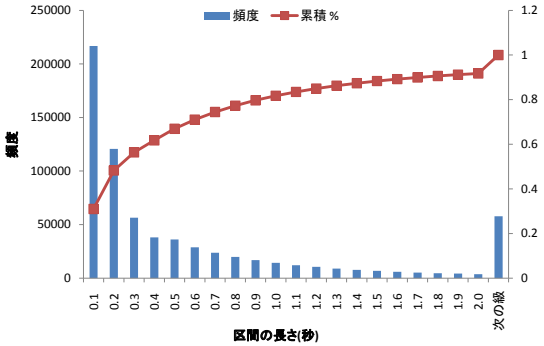

図 1 無音区間の長さ (0.1 秒単位) の分布
Fig. 1 Distribution of the lengths of silent intervals (unit length is 0.1sec)

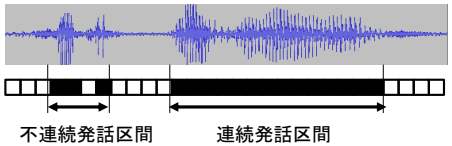

図 2 講義音声の発話区間への分割
Fig. 2 Sound segmentation into speech segments

ことなく講義の時間配分や進捗を把握することが考えられる．そこでこの節では，90 分の対面講義で収録した計 258 コマの講義音声を対象に,「1 秒以上の連続発話区間」「1 秒以上の不連続発話区間」を検出し，講義における最初（最後) の発話位置を求めることで講義開始・終了時の状況の分析を試みる．

講義音声の冒頭には「では」などの短い発話やマイクのスイッチを入れる音，教卓に物を置く音などの雑音が含まれているため，講義音声に最初に変化が現れた箇所は講義の開始位置には対応しない．通常，講義の初めには講義開始の合図となる言葉 (例：「みなさん．おはようございます.」) が発せられることから，ある程度の長さを持った不連続/連続発話区間を検出することで，実質的な講義の開始位置を検出できると期待される．

講義音声において最初/最後に 1 秒以上の長さを持つ連続発話/不連続発話区間が出現する位置を求め，各教員ごとに平均開始位置を求めた結果を表 2 に示す．この表から分かるように，連続発話区間と不連続発話区間の出現位置は一致しない．例えば，講義の冒頭に短い発話 (例：「えー」「では」「こんにちは」等) を伴う講義では，まず不連続発話区間が出現し，その後，本題に入る時に連続発話区間が出現することになる．

表 2 講師毎の発話区間出現位置集計結果
Table 2 Summary of the detected first/last speech segments

	不連続発話開始(秒)			連続発話開始(秒)			連続発話終了(秒)			不連続発話終了(秒)		
	平均(秒)	最小(秒)	最大(秒)	平均(秒)	最小(秒)	最大(秒)	平均(秒)	最小(秒)	最大(秒)	平均(秒)	最小(秒)	最大(秒)
講師1	28.9	0.0	125.1	41.2	0.0	125.1	5142.3	4261.7	5409.6	5184.2	4261.7	5651.4
講師2	82.0	13.3	217.4	115.8	13.3	231.6	5488.9	5358.3	5597.3	5495.7	5406.4	5597.3
講師3	53.7	0.0	163.8	67.5	0.0	187.9	5050.4	2074.9	5432.0	5063.9	2074.9	5432.0
講師4	12.9	0.0	26.3	16.8	0.0	50.9	3966.0	976.2	5018.2	3977.0	976.2	5032.8
講師5	80.1	0.0	385.2	516.6	24.1	3828.6	5115.4	2727.2	5678.9	5167.4	2824.5	5678.9
講師6	41.2	0.0	163.6	45.1	0.0	163.6	4054.0	711.2	5403.5	4316.1	711.2	5403.5
講師7	11.9	0.0	25.2	13.6	0.0	25.2	5214.3	4804.2	5381.0	5240.4	4804.2	5381.0
講師8	193.9	0.0	416.4	214.4	0.0	445.5	5183.6	3944.6	5389.0	5187.5	3944.6	5389.0
講師9	50.5	0.0	278.7	52.7	0.0	278.7	4642.4	2073.9	5388.0	4644.9	2079.7	5388.5
講師10	112.6	22.4	202.8	112.6	22.4	202.8	4500.8	4002.0	4999.5	4502.0	4004.4	4999.5
講師11	8.6	8.6	8.6	48.6	48.6	48.6	5490.3	5490.3	5490.3	5492.5	5492.5	5492.5
講師12	110.9	0.0	1229.9	111.5	0.0	1229.9	4848.3	2079.3	5746.5	4980.1	2079.3	5746.5
講師13	823.6	0.0	4394.2	855.3	0.0	4401.1	4640.7	1189.4	5688.8	4673.9	1189.4	5688.8
講師14	90.0	22.1	157.8	185.9	42.3	329.4	5045.0	4698.8	5391.2	5051.8	4698.8	5404.8
講師15	48.6	0.0	1033.5	54.0	0.0	1113.2	4762.8	1396.0	5556.4	4859.8	1396.0	5566.2
講師16	208.8	0.0	2204.7	213.5	0.0	2207.0	4843.8	2668.6	5603.9	4867.6	2668.6	5657.2
講師17	521.2	0.0	2123.4	558.2	0.0	2123.4	5225.6	3226.2	5691.3	5240.7	3226.2	5691.3
講師18	83.7	1.1	610.5	99.4	3.7	613.8	4672.2	976.2	5430.9	4754.6	1002.7	5496.4
平均	142.4	3.8	764.8	184.6	8.6	978.1	4882.6	2925.5	5460.9	4927.8	2935.6	5483.2

4.1 講義開始時点の発話区間分析

表 2 から分かるように講師によって発話開始時刻には講義開始数秒から 800 秒までばらつきがあるが，全講師の平均位置で見ると不連続発話の開始位置が 142 秒，連続発話の開始位置が平均 184 秒となっており，日常の講義実施状況から予想される値と大きな食い違いはない．講師によっては発話開始位置の最大値が 1000 秒以降と大きな値になっているが，これは音声収録の不具合によるものである (詳細は 4.3 節で後述)．

各講義音声に対して不連続発話開始位置と連続発話開始位置の差を求め，度数分布に直したのが図 3 である．この図から分かるように，60%以上の講義では不連続発話区間と連続発話区間が最初に出現する時刻に差はなく，時刻差が 30 秒以下のもので 90%を占める．

連続発話区間と不連続発話区間の最初の出現時刻が 10 秒以上異なるビデオを対象に講義状況を確認し，実際に何が起きているかの分析を行った．

10 秒ほどの時間差があったある講義の例では，講義の冒頭に「さて，はじめましょうか」と述べた後，前の方に座っている学生にプリントを配布し，10 秒ほど経ってから「えと，…」と改めて講義を開始している．同じ程度の時間差があった他の例では，最初にマイクを ON にしたときの音が検知され，その後，ピンマイクをネクタイにとりつける身支度をした後，「それでは始めましょう」と講義を開始していた．また,「あーあーあー」とマイクテストを

時刻差(秒)	頻度	累積 %
0	159	61.63%
5	26	71.71%
15	30	83.33%
30	18	90.31%
60	11	94.57%
120	7	97.29%
240	3	98.45%
240>	4	100.00%

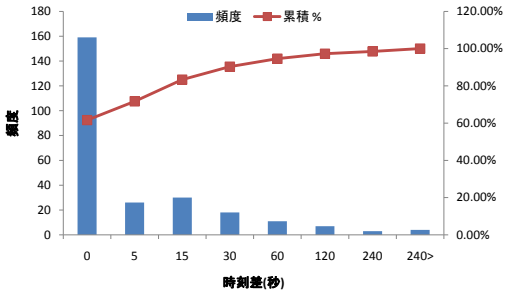

図 3 最初に出現する連続発話区間と不連続発話区間の時刻差の分布
Fig. 3 Distribution of the time differences between the first continuous and non-continuous speech segments

し，それから改めて講義を開始しているケースもあった．

数分以上の時間差があったケースでは，最初にまず講師マイクを ON にした音が検知され，それからプロジェクター・講義用 PC の起動をしているケースや，講義を始めたもののプロジェクターがうまく写らないためにパソコンやプロジェクターの設定を確認しているケースが確認された．

不連続発話区間と連続発話区間の出現位置が 4 分以上ずれていたケースが 3 つあり，うち 1 つは最初にマイクの電源を入れた後にプリント配布・レポート返却を行っているケースであった．残りの 2 件は講師がマイクのスイッチを入れ忘れたために収録機器の定常ノイズだけが収録されたケースであった．

以上は，不連続発話区間・連続発話区間の位置の違いが講義の進行の区切りを反映していたケースであるが，必ずしも発話区間の位置の違いが講義の進行の区切りに対応するわけではない．例えば，元々短い間を入れながら話すタイプの講師では 1 秒以上の連続発話区間が出現することが少ないため，講義の冒頭から解説をずっと話していても，不連続発話区間と連続発話区間の最初の出現位置が離れることがあった．また，講義の冒頭で出席を取るために学生の名前を点呼する講義でも同様の傾向があった．

4.2 講義終了時点の発話区間分析

表 2 における最後の発話区間終了時刻の平均を見ると，平均発話終了時刻には講師によって講義開始後 4000 秒から 5200 秒まで 20 分程度のばらつきがあることが分かる．90 分の講義においてこれだけの差が生じるのは，講師によって講義のスタイルが異なるためである．講義終了時刻をより詳しく分析するため，最後に検出された不連続発話区間の終了時刻をヒストグラムに直したものが図 4 である．7 割以上の講義において発話終了時刻が 5400 秒以降になっているが，残りの 2 割の講義では発話終了時刻が 60 分 ~90 分に分布している．該当科目と照らし合わせると，多くは講義時間の後半に演習を取り入れている科目であり，演習に割り当てる時間によって発話区間の終了時刻に違いが出ている．

終了時刻(秒)	頻度	累積 %
600	0	0.00%
1200	5	1.94%
1800	3	3.10%
2400	7	5.81%
3000	6	8.14%
3600	11	12.40%
4200	18	19.38%
4800	25	29.07%
5400	108	70.93%
6000	75	100.00%
次の級	0	100.00%

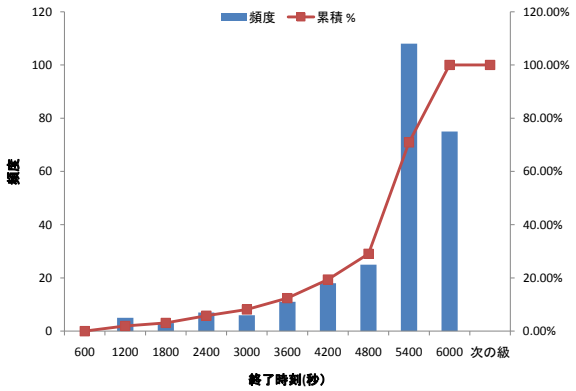

図 4 最後に出現する不連続発話区間の終了時刻の分布
Fig. 4 Distribution of the last moment of the last non-continuous speech segment

各講義音声に対して不連続発話終了位置と連続発話終了位置の差を求め，度数分布に直したのが図 5 である．この図から分かるように，50%以上の講義では不連続発話区間と連続発話区間が最後に出現する時刻に差はなく，時刻差が 30 秒以下のもので 80%を占める．時刻差が 5 秒未満のものよりも，時刻差が 5 秒以上 30 秒未満のもののほうが件数が多いことも分かる．

連続発話区間と不連続発話区間の最後の出現時刻が 10 秒以上異なるビデオを対象に講義状況を確認し，実際に何が起きているかの分析を行った．

出現時刻の差が 10 秒程度のケースでは，講義の最後の説明の中に短い間を入れて話しているために 1 秒以上の連続発話区間が検出されていないケースであった．

30 秒程度の時間差があったケースでは,「講義を終わりにしましょう」などの締めくくりの言葉を述べた後，すこし間をおいてマイクの電源を切るときのノイズ音が検出されていた．

30 分以上の差があったケースでは，講義の中盤以降マイクを OFF にしたままのため，ノ

時刻差(秒)	頻度	累積 %
0	140	54.26%
5	26	64.34%
30	49	83.33%
120	21	91.47%
240	12	96.12%
480	3	97.29%
960	2	98.06%
>960	5	100.00%

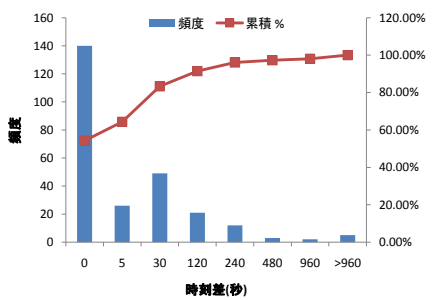

図 5 最後に出現する連続発話区間と不連続発話区間の時刻差の分布
Fig. 5 Distribution of the time differences between the last moment of continuous and non-continuous speech segments

イズ音が機械的に検出されていた．また，マイクの電池が切れかけのために講師の音声が途切れ途切れになり，途中から 1 秒以上の連続発話が検知できなくなったケースもあった．

全体として，最後にマイクの電源を切る時の音が最後の発話区間として検知されているケースが多く，単純に音声の振幅にもとづいて発話区間を検知するだけでは不十分であることが分かった．周波数解析を行うなど他の手法も検討する必要がある．

4.3 発話区間の開始・終了位置が特異な特徴を示した講義の例

講義によっては他の科目と発話区間の開始・終了時刻に大きな違いが生じるケースもあった．発話開始位置が講義開始時刻の 10 分以降だったケースを表 3, 発話終了位置が講義終了時刻の 25 分以前だったケースを表 4 に示す．

表 3 の講義 2, 講義 3, 講義 4 のケースでは，発話開始時刻が 14 分前後と他の科目に比べて遅い．このケースはいずれも同一の教員の同一科目であり，実際の講義状況をビデオで確認したところ講義の冒頭に演習を行ってから解説に入るというスタイルで実施していることが分かった．発話区間の検出位置と講義の実質的な開始位置がよく合致していると言える．

講義 8 のケースでは発話開始時刻が 20 分となっており，実際の講義状況をビデオで確認したところ，冒頭に小テストを行っていることが分かった．

表 3 に挙げた他の例はいずれもマイクの電源入れ忘れのために講義前半 (あるいは講義全体) の音声が収録できていないケースであった．

次に，講義終了時刻が早いケースについて実際の講義状況をビデオで確認した．表 4 の講義 15, 講義 16, 講義 18, 講義 19 のケースでは講義時間中に演習を取り入れており，講師による説明時間を意図的に少なくしている講義に該当した．講義 17, 講義 20 のケースでは講師マイクの電池が切れたために，講義の序盤以降の発話区間が検知できなくなっていることが判明した．

表 3 検出された発話開始時刻が 10 分目以降だったケース
Table 3 Lectures where first speeches were detected after 10-minute of their openings

	先頭発話区間検出時刻(秒)		
講義番号	不連続発話	連続発話	講義状況
講義1	687.1	963.2	マイクの付け忘れ→途中でON→無言で演習問題の解説を板書
講義2	837.7	851.7	演習見回り→「あ、あ」→身支度→「えっと、それでは・・・」
講義3	844.5	846.7	演習見回り→世間話→「はい、では・・・」
講義4	854.9	854.9	演習見回り→あいさつ→講義開始
講義5	996.7	1024.9	マイク完全に入れ忘れ
講義6	1033.5	1113.2	マイク完全に入れ忘れ
講義7	1192.7	1192.7	マイク入れ忘れ。途中からON。
講義8	1229.9	1229.9	冒頭小テスト→マイクON→講義開始
講義9	1964.6	1975.6	マイク入れ忘れ。途中からON。
講義10	2123.4	2123.4	マイク入れ忘れ。途中からON。
講義11	2204.7	2207	マイク入れ忘れ。途中からON。
講義12	2326.6	2329.5	マイク入れ忘れ。途中からON。
講義13	3955.2	3957.7	マイク入れ忘れ。途中からON。
講義14	4394.2	4401.1	マイク入れ忘れ。途中からON。

表 4 検出された発話終了時刻が 25 分目以前だったケース
Table 4 Lectures where last speeches were ended before 25-minute of their closings

	先頭発話区間検出時刻(秒)		
講義番号	不連続発話	連続発話	講義状況
講義15	711.2	711.2	演習内容をまとめて解説した後、マイクOFFで演習
講義16	738	900.1	演習内容をまとめて解説した後、マイクOFFで演習
講義17	976.2	976.2	途中でマイク電池切れのために音声が途切れ途切れ
講義18	976.2	1002.7	演習内容をまとめて解説した後、マイクOFFで演習
講義19	1189.4	1189.4	講義中にマイク電池切れ
講義20	1477.6	1477.6	演習内容をまとめて解説した後、マイクOFFで演習

4.4 連続発話区間に基づく講義状況の集計

講義中の講師の発話の長さを調べることで，平均的な説明の長さや，説明が長引いた箇所の検出ができると考えられる．ここでは 1 つの不連続発話区間を 1 つの発話と見なし，講義状況の集計を行う．

典型的な発話の長さを調べるために，不連続発話区間の長さの分布を求めたのが図 6 である．この図から分かるように，90%の発話区間は長さが 4 秒以下であり，99%の発話区間は長さが 10 秒未満である．長さが 20 秒を超える発話区間は全体の 0.2%以下であり，特異な事例と言える．

数分以上の長さを持つ不連続発話区間を見つけることで講義中の説明の長い箇所を見つけられるのではないかと考え，長さ 3 分以上の不連続発話区間の出現状況を分析した．該当区間は今回分析対象とした音声中に 69 カ所存在したが，長時間マイクを OFF にした際の

長さ(秒)	頻度	累積 %
1	126055	47.19%
2	64649	71.39%
3	32668	83.62%
4	17624	90.21%
5	10107	94.00%
6	5837	96.18%
7	3583	97.52%
8	2099	98.31%
9	1280	98.79%
10	844	99.10%
11	525	99.30%
12	383	99.44%
13	254	99.54%
14	173	99.60%
15	155	99.66%
16	111	99.70%
17	85	99.74%
18	73	99.76%
19	49	99.78%
20	44	99.80%
次の級	539	100.00%

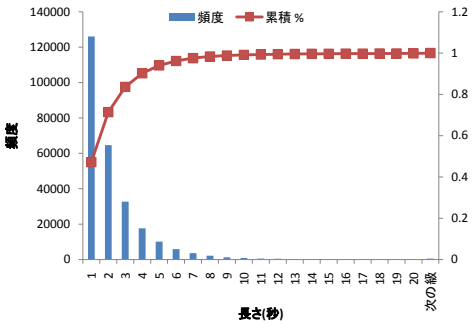

図 6 不連続発話区間の長さ (0.1 秒単位) の分布
Fig. 6 Distribution of the length of non-continuous speech segments (unit is 0.1sec)

講義	不連続発話区間	開始位置(分)	終了位置(分)	区間長(秒)	講義状況
講義1	区間1-1	0	6.1	363	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-2	6.1	10.3	248.7	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-3	10.3	14.3	244.4	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-4	16.2	21.3	303.5	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-5	25.1	34.1	535.3	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-6	34.1	46.9	768.8	板書・スライドで講義。話が区切れそうで区切れない。
	区間1-7	47.5	53.7	373.8	板書・スライドで講義。話が区切れそうで区切れない。
講義2	区間2-1	33.1	36.9	231.3	板書とスライドで計算過程を解説
	区間2-2	37	41.9	298.5	板書とスライドで計算過程を解説
	区間2-3	41.9	45.2	192.9	計算過程を板書とスライドで解説
	区間2-4	45.2	55	588	計算過程を板書とスライドで解説
	区間2-5	60.9	65.5	270.7	板書とスライドで計算過程を解説
	区間2-6	66.8	71.7	290.7	板書とスライドで計算過程を解説。67:30秒から無言で演習開始。黒板掃除の雑音が入る。
	区間2-7	71.7	78.7	424.3	73:40までは無音で演習見回り中。その後、演習問題の解説。
	区間2-8	81.3	94.9	810.9	途中からマイクON。終盤10分マイクOFF。
講義3	区間3-1	12.6	17	266.4	板書・スライドで講義。話が区切れそうで区切れない。
	区間3-2	20.8	25.1	258.6	板書・スライドで講義。話が区切れそうで区切れない。
	区間3-3	60.3	76	938.9	冒頭解説の後はマイクOFFのままで演習
	区間3-4	76.3	80.7	265.3	演習中。マイクOFFのまま

図 7 3 分以上の不連続発話区間を複数回含んでいた講義
Fig. 7 Lectures where non-continuous speech segments longer than three minutes were detected

定常ノイズが検出されているだけのケースが大半であった．しかしながら，実際に音声を聞いて定常ノイズに該当する区間を除外した結果，長い不連続発話区間を複数回含む特徴的な講義を見つけることができた (表 7)．

今回の集計では 0.6 秒以上の無音区間が出現する場所で機械的に発話区間を区切っているため，内容的にはひと続きの場面であっても，見かけ上複数の発話区間に分割されている．例えば表 7 中の講義 2 では実質的に区間 2-1〜 区間 2-4, 区間 2-6〜 区間 2-7 は時間的に連続しており，実際には 10 分以上連続して解説が行われていることが分かる．

5. 音声認識ソフトによる講義音声の分析

自動音声認識ソフトを講義音声に適用することで，発話内容の分析やビデオ字幕の書き起こしに利用することが考えられる．そこで今回はオープンソースとして広く利用可能な音声認識ソフト Julius を用い，収録音声の自動認識を試みた．

Julius では音響モデルと言語モデルを組み合わせて最も尤度の高い認識結果を出力するため，対面講義のように連続自然発話された音声を正しく認識するには言語モデルとして自然発話モデルを指定する必要がある．Julius で利用可能な自然発話モデルには国立国語研究所発行の「日本語話し言葉コーパス」があるが，2011 年 2 月下旬時点で在庫切れのため現在入手不能となっている．そのため，今回の分析では Julius ディクテーション実行キット[*1]付属の音響・言語モデル (Web から学習した語彙数 6 万のモデル) をそのまま用いている．

講義音声はディクテーション実行キット付属の音響モデルに適合するよう 16000Hz, モノラル形式に変換したものを用いた．また，Julius ディクテーション実行キットでは入力音声は 1 ファイルが 1 発話に対応するものと想定されているため，不連続発話区間に対応するように講義音声を分割し，分割後の各音声ファイルに対して音声認識を行っている．

5.1 音声認識ソフトによる講義音声の認識状況

講師音声認識結果の例を表 5 に示す．認識結果の多くは表中の発話 1〜 発話 7 のように発話内容と認識結果が大きく食い違っており，認識結果を見ても元の講義内容を把握することは不可能であった．しかし，話し方が明瞭な一部の講師については表中の発話 8〜 発話 11 のように，比較的良好な認識結果が得られることもあった．大半の講義音声については字幕書き起こしに使えるような品質の認識結果は得られず，対面講義用の音響モデル・言語モデルを適用することが不可欠と思われる．

なお今回の分析では一定以上の長さを持つ無音区間の出現位置で機械的に音声を分割しているため，音声の区切りと日本語の文としての区切りが必ずしも一致していない．音声の分割方法が認識精度に無視できない影響を与えた可能性も考えられる．

5.2 自動音声認識による発話冒頭語の分析

前節の結果から話し言葉に対する認識率が低いことが分かったが，発話回数の多い単語に関しては認識結果においても同様に多数出現することが期待される．特に，発話の話し始め

⋆1 今回の分析では Linux 版 Julius ディクテーション実行キット v4.1 を使用

表 5 Julius ディクテーションキットによる講師音声認識結果の例
Table 5 Examples of voice recognition results by Julius dictation kit

発話	原文	認識結果
発話1	ちょっと待ってね。電気消しとこう。	と なっ て ね ！ 元気 です 。
発話2	はい、前がこんだけ空いてんねん。	マイナス と なっ て ・ 。
発話3	えと、今日の資料はですね、えー、	四 仕様 で 、寝 て 。
発話4	えーと、この科目は、えー、三年生、を中心に	で 探せ 、五 、中心 に 。
発話5	情報の価値	じゃ この 感じ 。
発話6	(スー)皆さんは	N さん は ？
発話7	できた人から前に提出して終わりにしましょう。	形 です から 、満員 でし て 悔い 。
発話8	通信プロトコルを理解するっていうのが目的になる。	大変 感覚 を 理解 する って の が 目的 に なる
発話9	えー、後期が始まって今日が最初の授業かと思います。(スー)えーとー・・、ま、三年、	データ を 基 が 、始まっ て 、今日 が 最初 の 、事業 か と 思い ます 。デート 、馬鹿 な
発話10	今日の議題に入りますけど、	今日 の 時代 に 入り ます けど 。
発話11	えー、分担します。	で 、分担 し ます 。

の言葉をうまく検出することができれば，ビデオの編集の際のクリップ自動分割や，発話回数の分析に利用できると考えられる．

そこで，各講師の講義音声に対し，不連続発話区間の先頭部分がどのような音として認識されたかの集計を行った．各講義音声に対し，認識結果の先頭部分に出現した単語を出現回数の多いものから順に列挙したのが表 6 である．

表 6 から分かるように,「no」「kono」「de」のように全講師に対して検出回数が多い単語であっても，正解数が 0 の場合があるなど，検出回数の多い単語が必ずしも認識結果が正しいとは限らない．また，子音の認識誤りを許容し,「冒頭が一致する，母音の並びが一致する」ケースを「一部正解」と見なすと,「de」や「ne」の適合率が非常に高くなる (講師 2, 講師 4, 講師 5 の場合) が，必ずしも全ての講師に当てはまるわけではない．

講師による違いとしては，講師 2 の「kono」，講師 4 の「yuu」，講師 10 の「ano」のように 80%以上の高い精度で正しく認識される単語もある．その一方で，講師 1, 講師 8, 講師 11, 講師 12 のようにどの語の適合率も低いケースがある．

講義音声を自動分析することを考えると，認識誤りが少ない単語にはどのようなものがあるか興味がある．そこで，表 6 で分析したものと同じ講義音声に対し，検出結果がすべて正しかった単語を講師別にまとめたものが表 7 である．講師 12 名中，3 名については正解率 100%の単語が存在し，講義内容に関係するキーワードも検出されている．このことから，講師によって自動音声認識の適用のしやすさに違いがあり，特定の講師に対しては自動音声認識によりキーワードの抽出まで可能であると言える．

6. 考　　察

この節では以前の節で述べた内容にもとづき，講義音声の発話特徴を分析することで，講義ビデオのどのような応用が可能になるか考察する．

表 6 自動音声認識結果の発話冒頭語集計結果
Table 6 Voice recognition results of beginning parts of speech segments

			適合結果			適合率(%)		
講師	検出語	検出回数	正解	一部正解	不正解	正解	正解+一部正解	一部正解の例
講師1	no	27	0	0	27	0	0	
	kono	24	2	5	17	9	30	konpaira
	de	18	4	0	14	23	23	
講師2	de	27	3	19	5	12	82	e
	kono	15	12	0	3	80	80	
	jo:ho:	13	10	0	3	77	77	
講師3	ga	15	2	1	12	14	20	ma
	no	14	1	4	9	8	36	sono,kono,uchino
	de	14	7	2	5	50	65	e:
講師4	de	46	7	30	9	16	81	e:(28件)
	kono	28	14	10	4	50	86	konna(5件)
	yuu	26	25	0	1	97	97	
講師5	ne	30	1	24	5	4	84	de(22件)
	de	22	13	4	5	60	78	e:(4件)
	neqto	14	1	2	11	8	22	e:, e:to, de
講師6	de	13	5	6	2	39	85	e:, e:to
	desu	10	2	5	0	20	70	de, de:
	no	8	0	2	6	0	25	kono,sono
講師7	de	21	7	4	10	34	53	e:(3件)
	kono	16	6	8	2	38	88	kokono
	kore	15	9	3	3	60	80	kokoni(3件)
講師8	desu	15	2	3	10	14	34	de, desoreno
	no	14	0	2	12	0	15	mo
	de	14	5	4	5	36	65	e:(4件)
講師9	jibuN	15	6	0	9	40	40	
	desu	14	2	3	9	15	36	e:todesune
	kono	13	1	4	8	8	39	konnnamonde
講師10	kono	19	4	7	8	22	58	ano, sono
	ano	14	13	1	0	93	100	hannosuru
	to	10	0	0	10	0	0	
講師11	no	37	1	5	31	3	17	kono,sono
	ta	29	0	0	29	0	0	
	de	28	9	2	17	33	40	soide
講師12	de	24	5	2	17	21	30	e:(2件)
	ga	20	3	0	17	15	15	
	kono	18	4	2	12	23	34	sokono, sono

表 7 検出結果が全て正しかった単語のリスト
Table 7 Words that were correctly dictated by Julius dictation kit

講師	検出語(カッコ内は検出回数)
講師2	情報セキュリティ(3),考えて(3),コンピュータ(3),この(3),お金(3),そして(3),という(3),対策(6),それ(7)
講師3	今(4),データベース(11)
講師4	はじめ(3),モデル(3),最初(3),1980(3),正確に(3),大量のデータ(3),いうことで(3),こんな(4),なに(4),今(5),いうの(5),パソコン(6),コンピュータ(8)

4 節で見たように，発話区間の長さ・出現位置に特異な特徴を持つ講義を機械的に検出することが可能である．これにより，講義の冒頭・末尾に時間内演習を取り入れている科目や，講師が長時間説明を行っている場面を推定し，配信用ビデオの編集に反映させることが考えられる．また，講義の発話開始・終了位置が明らかに通常と異なるケースは結果としてマイクの不具合・利用忘れに対応しており，講義室の音響設備の健全性の確認や，音響機材の利

用説明対象者のリストアップといった業務にも応用が可能と思われる．

発話区間の判定においては，今回用いた判定方法では無発話時のマイクノイズの誤検出が多く，実際の収録用音響設備の特性に合わせた発話区間判定手法の開発が不可欠といえる．今回はマイク電源 ON/OFF 時のノイズを発話区間と誤認識するケースが多かったが，マイクノイズであることの判定を加えることで逆にマイクの利用開始・終了時刻の検知に利用することも考えられる．

5 節で見たように，自動音声認識については少なくとも自然発話モデルを適用しない素朴な状態では，字幕書き起こしやキーワード抽出に使える品質の情報は得られない．むしろ，特定の講師に対しては少ないながらも良好な認識結果が得られたことから，講師の発話の明瞭さを識別する指標として自動音声認識の結果を利用することが考えられる．発話の冒頭語については多くの講師について「de」「de:」「e:」が比較的精度良く検出できていたことから，これらの音に特化した音声検出手法を開発することで，発話開始位置を単位としてビデオの編集や配信を行うことが考えられる．

講義ビデオの長期収録によってこのような分析が可能になることについて，現場の講師からは以下のようなコメントがあった．

- 「自動的にキーワードを抽出し映像にラベル付けができれば，復習の際の効率が上がるのではないか」
- 「ミニ演習に割く時間など，授業時間の使い方を反省する際の材料になるかもしれない」
- 「学生に対して行う満足度アンケートの結果と照合することにより，どのような話し方をすると学生の理解度が深まるのかを考える材料にできるのではないか」
- 「音声認識の結果を参考にすることで，聞きやすい講義にする材料になるのではないか．より聞きやすい講義にするためにどう発音すればいいか，発話のクセの修正に活用できるのではないか.」

ビデオの分析を教育改善につなげることへの期待が多く，分析結果を教員にフィードバックするための仕掛けが重要になると思われる．

7. ま と め

講義ビデオを長期間大量に収録して発話特徴を分析することで，時間内演習を取り入れている講義や講師マイクの利用状況，発話の明瞭な講師の識別を行えることを見た．今回の分析方法は主として発話区間の検出にもとづいており，発話内容や講師の身振り，板書・スライド内容などの高次な情報を用いずに特徴的な講義や講義場面を検出できる．このため，収録映像の画質やマイクの音質にあまり左右されず幅広く適用可能であると考えられる．

今後の課題としては，発話区間の判定精度の向上，自然発話言語モデルの音声認識への適用，ビデオ編集・配信システムとの連携，分析作業の自動化などが挙げられる．

謝辞 本研究は科研費（課題番号:21700818）の助成を受けたものである．

参 考 文 献

1) Nagai, T.: Automated lecture recording system with AVCHD camcorder and microserver, *Proceedings of the ACM SIGUCCS fall conference on User services conference*, St. Louis, Missouri, USA, ACM, pp.47–54 (2009).
2) 永井孝幸：市販ハイビジョンカメラを用いた講義ビデオ撮影加工システムの運用報告，情報処理学会研究報告 第 1 回 CLE 研究発表会，Vol.2010-CLE-1, 情報処理学会，pp. 1–8 (2010).
3) 永井孝幸：鳥取-熊本間での講義ビデオ遠隔自動収録の試みについて，情報処理学会研究報告 第 3 回 CLE 研究発表会，Vol.2010-CLE-3, 情報処理学会，pp.1–8 (2010).
4) 土屋雅稔，小暮　悟，西崎博光，太田健吾，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作成と分析，情報処理学会論文誌， Vol. 50, No. 2, pp. 448–459 (2009-02-15).
5) 宗宮充宏，西崎博光，関口芳廣：E-024 話し言葉音声中のフィラー検出精度と音声認識率の関連性 (自然言語・音声・音楽, 一般論文)，情報科学技術フォーラム講演論文集, Vol.7, No.2, pp.191–192 (2008-08-20).
6) 西村雅史，伊東伸泰：講義コーパスを用いた自由発話の大語彙連続音声認識 (音声情報処理：現状と将来技術論文特集)，電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理， Vol.83, No.11, pp.2473–2480 (2000-11-25).
7) 西崎博光，関口芳廣：教員の話し方改善支援システムの開発に向けた講義音声の特徴分析 (< 特集 > 学習・教育支援のための技術開発)，日本教育工学会論文誌， Vol.34, No.3, pp.171–179 (2010-12-01).
8) 山口真之介，有馬明日菜，大西淑雅，西野和典，小林史典：音声認識ソフトウェアを活用した講義アーカイブシステムの検討，平成 22 年度　情報教育研究集会講演論文集 (2010-12-21).
9) 永井孝幸：有人撮影講義ビデオの閲覧・編集支援のための画像・音声切り出し手法の検討，情報処理学会研究報告 第 7 回 CMS 研究発表会，Vol.2007-CMS-7, 情報処理学会，pp.16–23 (2007).