論　文

ビデオ教材作成支援を目的とした講義音声によるシーン分割

金寺　登[†a)]　隅田　飛鳥[†]　池端　孝夫[†]　船田　哲男[††b)]

Subtopic Segmentation in the Lecture Speech for Creation of Lecture Video Contents

Noboru KANEDERA[†a)], Asuka SUMIDA[†], Takao IKEHATA[†], and Tetsuo FUNADA[††b)]

あらまし　ネットワーク上で利用できるビデオ教材は増えてきつつあるが，まだ少ない．ビデオ教材が少ない原因の一つはビデオ編集に手間と時間を要するためだと考えられる．そこで本論文ではビデオ教材作成支援を目的とし，編集前の講義ビデオの音声情報から自動的にシーン分割位置を推定する方法について検討する．ビデオの音声情報から認識を行った結果得られたテキスト情報より独立成分分析を用いて求められた指標を動的計画法により順次比較することでシーン分割位置推定を行った．5 人の教員による編集前の講義ビデオを用いて実験を行った結果，提案手法は Hearst 法と同等以上の分割性能をもちながら，分割数を自由に設定できることが分かった．また，音声認識結果を用いたシーン分割性能は書き起こしテキストと同等であることが確認された．

キーワード　ビデオセグメンテーション，ビデオ教材，独立成分分析，音声認識

1. まえがき

近年，高速なネットワーク環境が整備され，ビデオ教材を用いて自宅で手軽に予習・復習することが可能となってきた．しかし，利用できるビデオ教材はまだ少ない．この原因の一つはビデオを編集するために非常に多くの労力と時間を要するためだと考えられる．ビデオの編集作業にはビデオの取込み，シーン分割位置の検索，シーンの削除・移動・マージなどがある．特にシーン分割位置を検索するにはビデオを始めから最後まで繰り返し見る必要があり時間・労力ともに大きな負担となる．そこで，ビデオ教材作成を支援する方法として，ビデオシーンを自動分割するシステムの開発を検討している（図 1）．

自動的にシーン分割位置推定を行うために，ビデオ中の映像あるいは音声を用いる方法が研究されている．ビデオ中の映像を用いて自動シーン分割位置推定を行う研究に関して数多くの報告がある．これらの報告によれば，シーンの切換わり位置で映像が大きく変化する場合には高精度に分割を行うことができる [1]．しかし，講義ビデオの内容が映像と密接に関連して変化することは少ない．これに対し講義ビデオの内容は音声と非常に密接に関連して変化する．そこで本研究では音声を利用して編集前の講義ビデオのシーン分割を試みる．

音声を利用したシーン分割に関しては，講演 [2], [3] や編集後の講義 [4], [5] を対象としたものが多く，編集前の講義を対象としたものはほとんどない．編集前の講義をシーン分割する場合には不要文が数多く含まれるためトピックの切換わりの検出が更に困難であると予想される．そこで，本研究では編集前の講義ビデオ中から抽出された音声情報よりどの程度のシーン分割性能が得られるかを調査することを目的とする．

ビデオシーンの分割を行うには，テキストを指標（トピック表現）に変換する必要がある．指標には **2.** で述べるように様々な指標が提案されている．本研究では指標のサイズを小さくでき，人間にとって分かりやすい意味表現が得られる独立成分分析（ICA）を採用した．

シーン分割方法には **3.** で述べるように事前知識を必要とする方法と必要としない方法に大別される．ま

† 石川工業高等専門学校，石川県
Ishikawa National College of Technology, Ishikawa-ken, 929–0392 Japan
†† 金沢大学大学院自然科学研究科，金沢市
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa-shi, 920–8667 Japan
a) E-mail: kane@ishikawa-nct.ac.jp
b) E-mail: funada@t.kanazawa-u.ac.jp

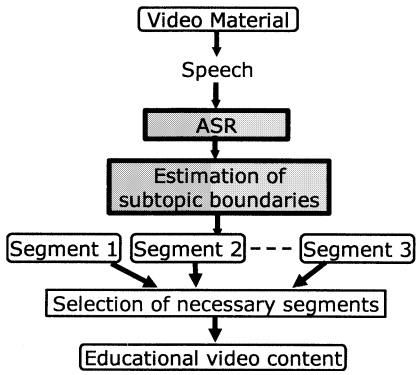

図 1 ビデオ教材作成支援システム例
Fig. 1 A video-segmentation system supporting the creation of lecture video material.

た，シーン内の話題の変動が小さくなるようにシーンを分割する方法 [6] とシーン間の話題の違いが大きくなる方法 [7], [8] に分類できる．本研究では事前知識を必要とせず，すべての隣接区間についての指標同士の cos（余弦）が最小になるように動的計画法を用いて最適な分割を求める方法を提案する．

2. シーン分割位置推定のための指標

2.1 シーンを指標へ変換する方法

一般的にビデオは複数のシーンからなっている．複数のシーンのうち，隣接するシーン間が似ていればひと続きのシーン，似ていなければそのシーン間にシーン分割点が存在すると考えることができる．シーン間をコンピュータ上で比較するためには，各シーンの話題情報を何らかの指標に変換しなければならない．指標には TF-IDF [2], [4] や TF-IDF を考慮した相互情報量 [5]，χ^2 値 [9] などがよく用いられている．また，語と文書行列の特異値分解（Latent Semantic Analysis；LSA）[10] や共起行列の特異値分解 [11] を用いてテキストを次元数が語彙数に依存しない指標に変換する方法が提案されている．更に Kabán ら [12], [13] は LSA の結果を独立成分分析（Independent Component Analysis；ICA）を用いて，人間に分かりやすい意味表現に変換する方法を提案している．

本論文ではビデオ音声を音声認識したテキストを，独立成分分析を用いて指標に変換し，シーン分割のための指標として応用する．シーンを指標に変換するには，まずビデオの音声情報から音声認識を用いてテキストに変換する．次に形態素解析を行い単語区切りにする．形態素解析器には茶筅 [14] を用いた．更に自立語のみを抽出し指標に変換する．

2.2 独立成分分析を用いたシーン分割のための指標

独立成分分析とは一般的に複数の独立した信号が混在した信号をもとの独立した信号に復元する方法をいう [15].

Kabán らは独立成分分析を用いて，語–文書行列（各文書における語の頻度）を，話題と語の関係，話題と文書の関係に変換する方法を提案した [12], [13]．この方法について簡単に説明する．

1. 独立成分分析の入力として，語–文書行列 D（T 行 N 列）を与える．ここで T は語数，N は文書数を表す．得たい話題数を K として語–文書行列 D を話題と語の関係を示す行列 $S^{(1)}$（K 行 T 列）と話題と文書の関係を示す行列 $S^{(2)}$（K 行 N 列）に変換することを考える．

2. D の特異値分解を次式で近似する．

$$D \cong UEV^T$$

ただし U は左特異値ベクトル（T 行 K 列），E は K 個の特異値をもつ正方行列（K 行 K 列），V は右特異値ベクトル（N 行 K 列）である．

3. $S^{(1)}$ がもとの独立した情報（話題）と語の関係を表す行列であるとする．この $S^{(1)}$ を変換行列 W で変換したものを $X^{(1)} = U^T$ と仮定する．また $S^{(2)}$ がもとの独立した情報（話題）と文書の関係を表す行列であるとする．この $S^{(2)}$ を変換行列 W で変換したものを $X^{(2)} = V^T$ と仮定する．$S = [S^{(1)}, S^{(2)}]$, $X = [X^{(1)}, X^{(2)}]$ としてまとめると次式のようになる．

$$S = W^T X$$

4. 一般的に独立成分分析では信号に非正規性を仮定し非正規性が最大になるような成分を求める [15]．ここでは非正規性を表す尺度としてフィッシャーのひずみ度（Fisher skewness）を用いる．フィッシャーのひずみ度が最大になるように S の各列の 3 乗の和を W について最大化すると式 (1) が得られる．よって次式を収束するまで繰り返すことにより S 及び W を求められる．

$$S = W_{old}^T X$$
$$W = X(S \circ S)^T \qquad (1)$$

ただし $\circ$ は要素ごとの積を表す．

$$W_{new} = W\sqrt{(W^T W)^{-1}}$$

TF-IDF による指標の要素数は語数 T × 文書数 N になる．一般的に語数 T は非常に大きいため，この要素数は非常に大きくなってしまう．TF-IDF に比べ，独立成分分析による指標の要素数は LSA と同様に話題数 K × 文書数 N であるため，指標のサイズを小さくすることができる．

独立成分分析による方法は，LSA を基盤としているため，性能的には LSA と同等と考えられる．しかし，独立成分分析による話題と文書の関係は各話題内の分散が最小となるクラスタリング結果と一致することが示されており [12]，LSA と比べ人間にとって理解しやすい指標である．

本研究では独立成分分析により求められた話題と文書の関係を表す行列 $S^{(2)}$ を指標 I として用いた．

3. シーン分割方法

編集前の講義ビデオを対象にシーン分割を行った例はないが，テキストセグメンテーションやニュースなどで使用されている方法が参考になる．これらの方法は事前知識を用いるものと事前知識を用いないものに分けられる．

事前知識として，講義や講演のビデオ素材とは別に配付資料などのテキストが用意できる場合には，ビデオ中の音声とテキストを対応づけることでシーンを効率的に分割できる [3], [4]．また，話題の転換点を表す談話標識を事前知識として利用することもできる [2]．更に単語の共起関係も事前知識として利用できる [6], [11]．ニュースの分割において各話題のテキストデータで HMM を事前に学習させることにより良好な分割性能が得られている [16]．

事前知識を用いないもの，すなわち学習が不要で入力テキストのみで分割可能な方法には同一シーン内の同一単語の繰返し数が最大になるように分割する方法 [17] や同一シーン内変動を最小にする方法 [6]，隣接シーン間を比較する方法 [5], [7], [8] が提案されている．

本論文では編集前の講義ビデオを対象に事前知識が与えられない場合を想定し，隣接シーン間を比較する方法を用いる．以下，隣接シーン間を比較する Hearst 法と提案方法について説明する．

3.1 Hearst 法によるシーン分割 [7], [8]

Hearst 法では隣接するブロック間における類似度の変化を用いてシーンを分割する．ここでブロックは一定の語数から構成される語の列を意味する．隣接するブロック間における類似度の変化はブロック間の境界をシフトしながら類似度を計算することで求められる．隣接するブロック間における類似度の変化のうち，極大値と極小値との差が大きい部分にシーン境界が存在すると想定し，極小値に対応するブロック間の境界位置を分割位置とする．

まず図 2 に示すように，単語 W_i より左側の一定語数のブロックを BL_i，右側のブロックを BR_i とする．このときブロック間の境界位置 i における，ブロック間の類似度 y_i を次式で定義する．

$$y_i = \frac{\sum_t w_{t,BL_i} w_{t,BR_i}}{\sqrt{\sum_t w_{t,BL_i}^2 \sum_t w_{t,BR_i}^2}} \tag{2}$$

ここで t は各語，$w_{t,BL}$ は単語 t がブロック BL に出現する頻度を表す．次に W 語分だけブロック間の境界位置をシフトして，同様に類似度を求める．これを文末まで繰り返し，類似度の変化を求める．類似度の極大点は話題が盛り上がっている部分，極小点はシーン境界と想定される．つまり次式で与えられる類似度の極大値と極小値との差が大きければ大きいほどシーン境界らしいといえる．

$$score(j) = (y_l - y_j) + (y_r - y_j) \tag{3}$$

ここで j は注目している極小点のブロック境界位置，l は j から左にある最初の極大点のブロック境界位置，r は j から右にある最初の極大点のブロック境界位置，y_l，y_j，y_r はそれぞれ l，j，r における類似度を表す．よって $score$ が大きい順から境界候補とみなすことができる．なお類似度の微弱な振動を除去するために $score$ を求める前に類似度に対してスムージングを行うことが多い．

3.2 DP によるシーン分割

図 3 にシーン分割方法の概要を示す．一般的に隣接

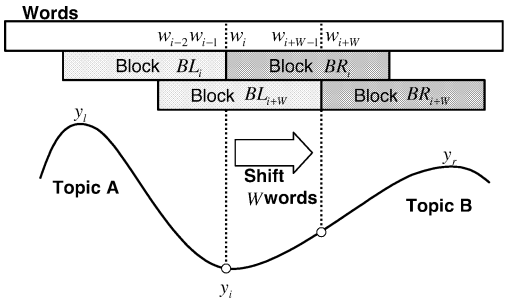

図 2 Hearst 法によるシーン分割
Fig. 2 Scene segmentaion by Hearst method.

するシーン間が似ていればシーン間がひと続きのシーンであると考えられ，似ていなければシーンが分割できると考えられる．このことから隣接するシーンの指標間における余弦の和が最小となるようにシーン分割すればよい．Hearst 法ではブロックが固定長であるため，ブロックは厳密にシーンに対応しているわけではない．また，分割数を指定することが難しい．

提案方法ではブロックを可変長とし，厳密にシーンに対応させる．また分割数を指定することができる．これによりビデオ編集システムにおいて，一部区間を詳細に分割したり，荒く分割したりすることが容易になる．以下に具体的な手順を示す．

まずビデオを仮にいくつかの文書区間に分割する．次に文書区間ごとに指標に変換し隣接する文書区間ごとに似ているかどうかを調べる．隣接する文書区間が似ているかどうか調べるには指標の余弦を用いる．余弦が小さい程文書区間は似ておらず，大きい程文書区間は似ていると考えられる．つまり，指標の余弦の総和が最小であればすべてのシーン間が似ていないことになり，文書全体を適切にシーン分割できると考えられる [5]．そこで，本論文ではシーン分割位置推定を余弦の総和が最小となるようなシーンの組合せを探す問題とみなし，動的計画法（Dynamic Programing；DP）を用い解く方法を提案する．

指標 I を用いて文書 1〜N を P 分割し 1〜b_1, (b_1+1)〜$b_2, \ldots, (b_{P-1}+1)$〜N の文書区間にすることを考える．指標 I の各列は文書 1〜N に対応する．p 番目の文書区間 $(b_{p-1}+1)$〜b_p に対応する I の和 $\boldsymbol{r}_p$ を次式で定義する．

$$\boldsymbol{r}_p = \sum_{m=b_{p-1}+1}^{b_p} I_m$$

ただし I_m は I の第 m 列とする．隣接する $\boldsymbol{r}_p$ と $\boldsymbol{r}_{p+1}$ の余弦の和が最小になるようにシーン境界 $\hat{B}_P = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{P-1})$ を次式で決定する．

$$\hat{B}_P = \arg\min_{B_P} \sum_{p=1}^{P-1} d(\boldsymbol{r}_p, \boldsymbol{r}_{p+1}) \tag{4}$$

$d(\boldsymbol{a}, \boldsymbol{b})$ はベクトル $\boldsymbol{a}$, $\boldsymbol{b}$ 間の余弦測度を表す．余弦測度を求める際，ベクトル $\boldsymbol{a}$ を求める際に使用した語数が一定の語数に満たない場合，$\boldsymbol{a}$ より前方に三角スムージングを行った．同様にベクトル $\boldsymbol{b}$ を求める際に使用した語数が一定の語数に満たない場合，$\boldsymbol{b}$ より後方に三角スムージングを行った．

三角スムージング方法を図 4 に示す．語数が一定の語数に満たない場合，一定の語数に達するまで前後の文書を追加し，指標を求める．ただしスムージングによる効果が大きくなりすぎないよう，追加した文書の重みを図 4 に示すように徐々に減少させる．

式 (4) を解くために動的計画法を用いる．まず，文書 1〜i を j 分割したときの隣接文書区間の累積余弦測度 $g(i,j)$ を次式で定義する．

$$g(i,j) = \min_{B_j} \sum_{p=1}^{j-1} d(\boldsymbol{r}_p, \boldsymbol{r}_{p+1}) \tag{5}$$

ただし $B_j = (b_1, b_2, \ldots, b_{j-1})$ である．

また，文書 k から i までの指標 $\boldsymbol{s}(k,i)$ を

$$\boldsymbol{s}(k,i) = \sum_{m=k}^{i} I_m$$

とすると，

$$\boldsymbol{r}_p = \boldsymbol{s}(b_{p-1}+1, b_p)$$

となり，以下のようにシーン境界を求めることができる．

1. $j=1$ のとき $i = 1, 2, \ldots, N$ について

$$g(i,1) = 0$$
$$b(i,1) = 0$$

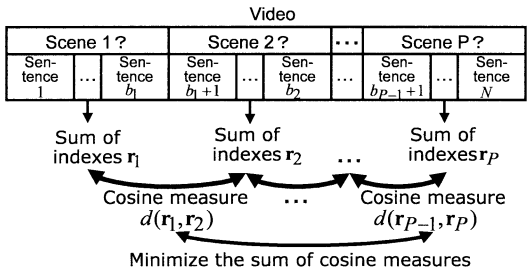

図 3　シーン分割方法概要
Fig. 3　Outline of scene segmentation method.

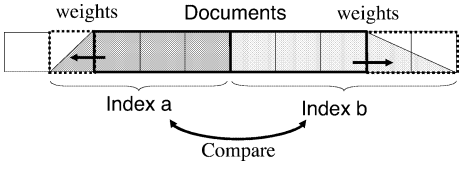

図 4　三角スムージング
Fig. 4　A triangular smoothing.

2. $j \geq 2$ のとき $i = j, j+1, \ldots, N$ について

$$b(i,j) = \mathop{\arg\min}_{k=j-1,\cdots,i-1} \big\{ g(k, j-1) + d\big(\boldsymbol{s}(b(k,j-1)+1, k), \boldsymbol{s}(k+1, i)\big)\big\}$$

$$g(i,j) = g(b(i,j), j-1) + d\big(\boldsymbol{s}(b(b(i,j),j-1)+1, b(i,j)), \boldsymbol{s}(b(i,j)+1, i)\big)$$

3. $\hat{b}_P = N$

$p = P-1, P-2, \ldots, 1$ について

$$\hat{b}_p = b(\hat{b}_{p+1}, p+1)$$

図 5 に文書 1〜i を 3 分割したときの隣接文書区間の累積余弦測度を求める例を示す．3 番目の文書区間が文書 $(k+1)$〜i であると仮定したとき，3 番目の文書区間の指標は $\boldsymbol{r}_3 = \boldsymbol{s}(k+1, i)$ で与えられる．2 番目の文書区間が k で終了するとすれば，2 番目の文書区間の開始点は $b(k,2)+1$ で与えられ，2 番目の文書区間の指標は $\boldsymbol{r}_2 = \boldsymbol{s}(b(k,2)+1, k)$ で与えられる．したがって 2 番目の文書区間が k で終了するとき，$g(i,3)$ は，文書 1 から k を 2 分割したときの累積余弦測度 $g(k,2)$ に，2 番目の文書区間の指標 $\boldsymbol{r}_2$ と 3 番目の文書区間の指標 $\boldsymbol{r}_3$ の余弦を加えたものになる．よって，すべての k について $g(k,2) + d(\boldsymbol{r}_2, \boldsymbol{r}_3)$ を求め，その最小値を $g(i,3)$ とすればよい．

Hearst 法では固定長の隣接ブロックを比較するのに対し，提案手法では可変長の隣接ブロックを比較するため提案手法の計算量は Heast 法に比べて，DP を用いても多くなる．

Hearst 法では，厳密に隣接シーン間の比較を行わず，固定長のブロックを比較することで計算量を $O(N)$ に抑えている．厳密に隣接シーン間の比較を行うには $O({}_N C_{P-1})$ の計算量が必要であるが，DP を用いることで提案手法では $O(N \times P)$ の計算量に軽減している．厳密に隣接シーン間の比較を行うことにより，ビデオ編集に必要な高い再現率が得られることを期待できる．

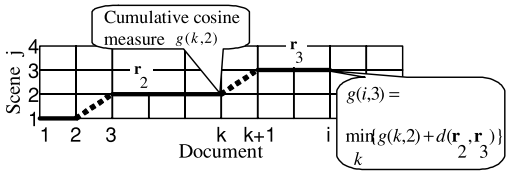

図 5　動的計画法によるシーン分割点探索例
Fig. 5　A scene segmentaion example by DP.

4. シーン分割結果

4.1 実験条件

実験対象として，表 1 に示すビデオ素材を用意した．これらのビデオ素材は，5 名の男性話者による約 90 分の講義 5 回分である．表 1 における文数は 1 s 以上の無音区間が継続するかどうかで区切られた境界候補（音声区間）数である．本実験ではこの一つの音声区間を **3.2** における 1 文書とする．収録には接話型ヘッドセットを用いたため，雑音等の影響は少ない．対象となるビデオ素材から音声情報のみを抽出し，16 kHz にダウンサンプリングを行った．次に音声区間ごとに日本語ディクテーション基本ソフトウェア（98 年度版）[18] を用いて音声認識を行った．音響モデルは 2000 状態 16 混合のトライフォンとし，各種学習・評価条件は文献 [18] と同様である．ただし，学習データには男性話者のみを用いた．言語モデルは講演の書き起こしテキストにより学習された言語モデル [19] を用いて認識を行った．認識結果から得られた文書の単語正解率・単語正解精度・未知語率を表 1 に示す．

音声認識によって得られたテキストから自立語のみを抽出後，提案手法・Hearst 法を用いて分割を行いシーン境界候補を求めた．必要以上に長い無音区間は不要部分として削除される可能性が高いため，5 s 以上の無音区間の両端もシーン境界候補に追加した．

提案手法では，独立成分分析（ICA）による指標を用いシーン分割を行った．独立成分数は，予備実験 [20] より表 1 に示されている文数の約 0.15 倍とした．ま

表 1　ビデオ素材
Table 1　Lecture video materials.

ビデオ素材	文数	共通正解境界数	単語正解率(注1)	単語正解精度(注2)	未知語率 [%]
1	539	21	50.9	33.5	7.1
2	592	23	46.7	31.2	2.9
3	544	14	40.3	22.7	8.2
4	468	18	32.0	12.4	5.5
5	430	24	45.8	26.3	8.6
平均	515	20	43.1	25.2	6.5

(注1)：単語正解率 [%] $= \dfrac{\text{総単語数} - \text{置換誤り単語数}}{\text{総単語数}} \times 100$

(注2)：単語正解精度 [%]
$= \dfrac{\text{総単語数} - \text{置換誤り単語数} - \text{付加単語数}}{\text{総単語数}} \times 100$

た，指標を求める際，100 語に満たない場合に限って三角スムージングを行った．

Hearst 法では窓の幅を 80 語とし，8 語ずつシフトしながら類似度を求めた．また前後二つの類似度を平均することによって類似度のスムージングを行った．

正解データとして，5 名の評価者に対象としたビデオ素材を提示しシーン境界の許容範囲を求めてもらった．許容範囲が 3 名以上一致する範囲を OR 合成し，正解とした．各データの共通正解境界数を表 1 に示す．

以下に評価尺度として用いた再現率，適合率の式を示す．

$$\text{再現率 } (recall) = \frac{\text{回答中の正解数}}{\text{正解数}}$$

$$\text{適合率 } (precision) = \frac{\text{回答中の正解数}}{\text{回答数}}$$

本研究では再現率を優先した．シーン分割位置推定において再現率が低い場合，ユーザが逐次的にシーン分割点を検索しなければならず，労力は大きいと考えられる．一方，余分に付加されたシーン分割点は無視すればよい．よって，シーン分割位置推定による誤りが少ない（再現率が高い）ことが望ましい．

4.2 実験結果

図 6 に音声認識結果によるシーン分割結果を示す．横軸は分割数を各ビデオ素材の文数で割った分割率で，縦軸は再現率若しくは適合率の平均を表している．この結果より独立成分分析（ICA）を用いた指標による場合，TF-IDF と同等以上の結果が得られることが確認できた．TF-IDF を用いて DP を行う際の指標の次元数は語数（一般的に大きな数）となるが，独立成分分析を用いた場合の次元数は指定した独立成分の数となるため，独立成分分析は語数にかかわらず TF-IDF に比べて高速にシーン分割が可能となる．語彙数は少なくとも 1000 単語以上であるため，独立成分数が 100 の場合，ICA の計算量及び記憶容量は TF-IDF に比べて 1/10 以下でよい．なお，独立成分分析を行う計算量は，隣接シーン間の比較を行う計算量に比べて無視できる程度である．

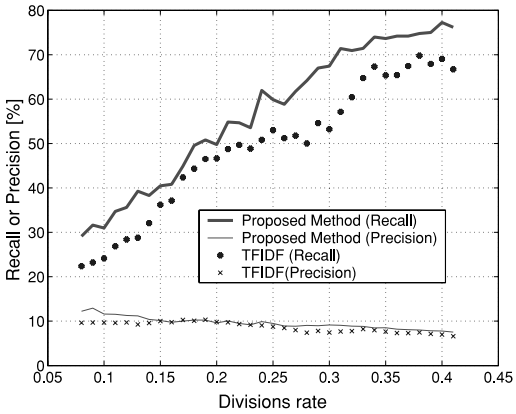

図 6 音声認識結果によるシーン分割結果（ICA と TF-IDF の比較）

Fig. 6 Scene segmentation results by automatic speech recognition (Comparison of ICA and TF-IDF).

図 6 はいずれも 5 s 以上の無音区間の両端もシーン境界候補に追加した結果である．5 s 以上の無音区間の検出のみによる再現率は 20.5%，適合率は 11.8%，分割数は全文数の 6%であった．

表 2 に Hearst 法を用いて音声認識結果よりシーン分割を行った結果を示す．Hearst 法では **3.1** における式 (3) の score がしきい値より大きい場合にシーン分割位置とする．しきい値を式 (2) の類似度の平均 − 標準偏差/2 と平均 − 標準偏差の場合について調査したが，ほとんど分割率（シーン分割数/文数）が変化しなかった．Hearst 法では式 (2) の類似度の極小値をシーン境界候補とするが，極小値の総数が限られているため分割数を極小値の総数以上に大きくすることができない．したがって Hearst 法ではビデオ編集に必要な高い再現率を得ることができない．一方，提案方法では図 6 に示すように自由に分割数を制御することができる．これにより，ビデオ編集を行うユーザの編集方針に応じて分割数を自由に設定することが可能である．また，Hearst 法（表 2）と提案手法（図 6）の比較より，提案手法は Hearst 法と同等以上の精度が得られていることが分かる．

提案手法を用い，音声認識結果と書き起こしテキストによるシーン分割を行った結果を図 7 に示す．この結果より音声認識結果を用いたシーン分割性能は書き起こしテキストと同程度であることが確認された．これは音声認識性能がある程度低くても複数個所において同じ誤りであればシーン分割には影響を与えないためと考えられる．複数個所において同じ誤りであれ

表 2 シーン分割結果（Hearst 法）

Table 2 Scene segmentation results (Hearst method).

しきい値	分割率	再現率 [%]	適合率 [%]
平均 − 標準偏差/2	0.09	28.4	12.4
平均 − 標準偏差	0.09	29.2	12.7

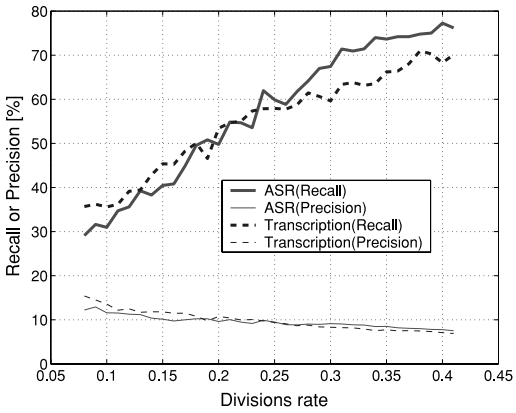

図 7　音声認識結果と書き起こしテキストによるシーン分割結果

Fig. 7　Scene segmentation results by automatic speech recognition and transcription text.

ば正解とした場合の実質的な単語正解率は 16.2%向上し，59.3%であった．

今回使用した評価データは，実際の講義を収録した話し言葉であるため，書き起こしテキスト自体にも言い誤りが多く含まれ，日本語の文法を満たさないことも多かった．また，新聞記事やニュースとは異なり，一つの講義には 1〜2 のトピックしか含まれないため，実質的にはサブトピック境界の検出となった．したがって，見かけ上，低い再現率，適合率となった．16 名の評価者による予備実験 [21] では，人間でも平均再現率が 40.5%であり，サブトピック境界は個人差が大きいと考えられる．更に編集実験では，16 名の被験者のうち，75%の被験者が提案方法で自動シーン分割を行った方が編集しやすいと回答した．

5. むすび

ビデオ教材作成支援を目的として，編集前の講義ビデオ中の音声情報により，ビデオシーンを自動分割した．ビデオシーンの分割には，独立成分分析を用いたトピック表現（指標）とポーズ情報を利用した．シーンの対応付けには DP を用い，隣接するシーンの余弦の総和が最小になるように最適化した．実験の結果，提案手法を用いることで，Hearst 法とほぼ同等以上の分割性能をもちながら，分割数を自由に設定できることが分かった．独立成分分析による指標を用いた場合，TF-IDF とほぼ同等以上の結果が高速に得られることが分かった．また，音声認識結果を用いても書き起こしテキストと同程度のシーン分割性能が得られることが確認された．

謝辞　本研究の一部は文部科学省科学研究費補助金（課題番号 14580246）を受けて行われた．

文　献

[1] 中村裕一，外村佳伸，“見たい部分を簡単に短時間で，” 信学誌，vol.82, no.4, pp.346–353, 1999.

[2] 長谷川将宏，秋田祐哉，河原達也，“談話標識の抽出に基づいた講演音声の自動インデキシング，” 情処学研報，2001-SLP-36-6, pp.35–42, 2001.

[3] 伊藤克亘，藤井　敦，石川徹也，“音声文書検索を用いたオンデマンド講義システム，” 信学技報，SP2001-111, 2001.

[4] 山本夏夫，緒方　淳，有木康雄，“トピックセグメンテーションに基づく講義ビデオの構造化の検討，” 情処学研報，2002-SLP-42-10, pp.59–64, 2002.

[5] 緒方　淳，山本夏夫，鷹尾誠一，有木康雄，“講義データを対象とした音声認識と構造化の検討，” 情処学研報，2001-SLP-37-14, pp.79–84, 2001.

[6] 別所克人，“クラスタ内変動最小アルゴリズムに基づくトピックセグメンテーション，” 情処学論（自然言語処理），vol.154, no.25, pp.177–183, 2003.

[7] M. Hearst, “Multi-paragraph segmentation of expository text,” 32nd. Annual Meeting of the Association for Computational Linguistics, pp.9–16, 1994.

[8] M. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” Computational Linguistics, vol.23, no.1, pp.33–64, 1997.

[9] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui, “Topic extraction based on continuous speech recognition in broadcast news speech,” IEICE Trans. Inf. & Syst., vol.E85-D, no.7, pp.1138–1144, July 2002.

[10] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Laundauer, and R. Harshman, “Indexing by latent semantic analysis,” J. Am. Soc. Inf. Sci., vol.41, no.6, pp.391–407, 1990.

[11] 別所克人，“単語の概念ベクトルを用いたテキストセグメンテーション，” 情処学論，vol.42, no.11, pp.2650–2662, 2001.

[12] A. Kabán, Latent Variable Models With Application to Text Based Document Representation, Ph.D. Thesis, The University of Paisley, 2001. http://cis.paisley.ac.uk/kaba-ci0/ata_thesis.zip

[13] A. kabán and M.A. Girolami, “Fast extraction of semantic features from a latent semantic indexed corpus,” Neural Process. Lett., vol.15, no.1, pp.31–43, 2002.

[14] 松本裕治，北内　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸，“日本語形態素解析システム『茶筌』version 2.2.1 使用説明書，” Technical Report, Nara Institute of Science and Technology, 2000. http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1-j.pdf

[15] 北　研二，津田和彦，獅子堀正幹，情報検索アルゴリズム，共立出版，2002. ISBN4-320-12036-1.

[16] J.P. Yamron, I. Carp, S. Lowe, and P. van Mulbregt,

"A hidden Markov model approach to text segmentation and event tracking," Proc. ICASSP-98, vol.1, pp.333–336, 1998.
[17] F.Y.Y. Choi, "Advances in domain independent linear text segmentation," Proc. NAACL-2000, 2000.
[18] 河原達也, 李　晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山田　篤, 宇津呂武仁, 鹿野清宏, "日本語ディクテーション基本ソフトウェア 98 年度版," 音響誌, vol.56, no.4, pp.255–259, 2000.
[19] 南條浩輝, 加藤一臣, 李　晃伸, 河原達也, "大規模な日本語話し言葉データベースを用いた講演音声認識," 信学論（D-II）, vol.J86-D-II, no.4, pp.450–459, April 2003.
[20] 隅田飛鳥, 金寺　登, 寺家谷純, 池端孝夫, 船田哲男, "独立成分分析を用いた音声による講義ビデオシーン分割," 信学技報, SP2003-61, 2003.
[21] 金寺　登, 池端孝夫, 隅田飛鳥, 船田哲男, "講義音声認識を用いたビデオ教材作成支援システムの評価," 音響学秋季講論, pp.37–38, 2004.

（平成 16 年 5 月 26 日受付, 11 月 22 日再受付）

金寺　　登　（正員）

昭 58 石川高専卒. 昭 60 電通大・通信卒. 昭 62 東大大学院修士課程了. 同年石川高専助手. 1996 Oregon Graduate Institute of Science and Technology（USA）客員研究員. 現在石川高専助教授. 博士（工学）. 音声認識の研究に従事. IEEE, 日本音響, 情報処理, 人工知能学会各会員.

隅田　飛鳥

平 15 石川高専・電子情報卒. 現在同校専攻科電子機械工学専攻に在学中. ビデオの自動セグメンテーションに関する研究に従事.

池端　孝夫

平 15 石川高専・電子情報卒. 現在同校専攻科電子機械工学専攻に在学中. ビデオ作成支援システムの構築に関する研究に従事.

船田　哲男　（正員）

昭 41 金沢大・工・電子卒. 昭 46 名大大学院博士課程了. 昭 46 金沢大・講師. 現在同大教授, 生体情報処理, 音声情報処理の研究に従事. 共著「数値解析の基礎」,「音声情報処理」など. IEEE, 日本音響学会, 日本エム・イー学会, 情報処理学会各会員.