CN-005

反転授業における音声明瞭性を考慮した講義映像作成支援システムの開発

Developing a Lecture Video Editor to Improve Voice Clarity for Flipped Classrooms

松浦辰雄 † 大囿忠親 † 新谷虎松 †

Tatsuo Matsuura Tadachika Ozono Toramatsu Shintani

1. はじめに

生徒の学習意欲の向上や知識の定着を促す新しい学習スタイルとして授業と宿題の役割を反転させる反転授業を取り入れる教育機関が増加している．反転授業では，生徒が自宅で学習するために教師があらかじめ講義映像を作成および準備を行う．撮影した映像をそのまま講義映像として配信することは少ない．映像の不要な部分を削除したり，生徒が講義内容を理解しやすくするために編集を行う必要があり，教師の映像の編集における負担は大きい．

そこで，本研究では，反転授業における講義映像の作成および編集の支援を行う．反転授業で用いられる講義映像の形式はスライドなどの静止画を用いて解説したり，板書や PC の作業画面を映し，書き込みながら解説する形式がある．本研究では，生徒の講義内容の理解に影響する解説音声に注目している．撮影した映像の音声の中で早口や口癖が出ている部分を検出し，半自動的に聞き取りやすい音声に編集を行う講義映像作成支援システムを開発した．ユーザは，本システムを用いることで生徒が聞き取りやすく，理解しやすい映像講義を効率良く作成および編集することができる．本稿では，講義映像の言葉の間，早口および口癖を半自動的に調整を行うシステムである講義映像作成支援システムの実装について述べる．

2. 講義映像の導入および編集

2.1 講義映像の導入

反転授業における講義映像の導入について述べる．授業と宿題の役割を反転させる反転授業は，生徒の学習意欲の向上や知識の定着を促すことを目的とし 2000 年頃から提案されている．生徒が自宅で講義映像を使用してあらかじめ学習をし，教室ではグループ学習やディスカッションなどの授業として発展した内容を行う．また，教育コンテンツとして講義映像を提供しているサービスも登場し，JMOOC[1]や Khan Academy[2]を利用することで反転授業の導入コストは低下している．しかし，教師の指導方針に合わせたり専門性の高い講義については教師自身が講義映像を作成する機会も多い．撮影した映像の不要な部分を削除したり，生徒が講義内容を理解しやすくするために編集を行うなど教師の講義映像の編集コストは依然高いままである．そこで，本研究では，反転授業における講義映像の作成および編集の支援を行い，教師の講義映像の編集コストの低減を目的とする．

また，反転授業で用いられる講義映像の形式は，スライド形式，作業画面形式および擬似講義形式が考えられる．スライド形式は，スライドなどの静止画を用いながら解説を行う．作業画面形式は，板書やパソコンの作業画面を映し，書き込みながらもしくは作業を行いながら解説を行う．擬似講義形式は，実際に教師が映り黒板を使って擬似的な講義を行う．特に，スライド形式および作業画面形式では，解説音声の質が講義内容の理解に大きく影響すると考えられる．本研究では，スライド形式および作業画面形式での解説音声に対して，生徒が講義映像を聞き取りやすくおよび理解しやすくするために半自動的に編集を行う．

2.2 音声明瞭性を考慮した編集

本研究では，音声明瞭性を講義映像の音声の聞き易さとし，撮影した映像の音声情報に対して半自動的に音声明瞭性の高い音声情報に編集を行う．また，本研究では，音声明瞭性を向上させるために講義映像の音声情報の言葉の間，発話速度および口癖に注目する．

言葉の間が詰まっていると生徒が講義内容に対して考えるゆとりがなくなるため言葉の間の調整を行う．言葉の間の長さや位置について須藤ら [1] によると，0.4 秒未満の短い言葉の間や発話の区切りとして不自然な位置で言葉の間がとられていたりすると聴講者に発話内容の理解に悪影響を与える．さらに，接続詞の後，名詞の列挙における体言止めおよび名詞＋助詞の後に言葉の間が挿入されることが自然な位置で取られる言葉の間であるとされる．本研究では，映像講義の音声情報が接続詞＋「間」，名詞＋「間」＋名詞および名詞＋助詞＋「間」で 0.5 秒取られているか判定し，言葉の間の調整を行う．

早口である音声は，生徒が講義内容を理解するのに弊害を与えるため早口の調整を行う．本研究では，一定の時間長を持つ音の文節単位であるモーラ (mora) を基準に早口の判定を行う．日本語の場合，仮名一つの単位が 1 モーラに当たり，小書き仮名は 1 モーラとは数えない．例えば,「しゃ」など仮名＋小書き仮名で 1 モーラと数える．本研究では，講義映像の音声において 1 秒当たりのモーラを算出し，音声情報の発話速度の変更を行い早口の調整を行う．また，発話速度の変更により声高が変化し，講義映像を視聴している生徒に違和感を与えてしまうことが考えられる．そこで，声高を変えずに音声の伸縮を行う手法を用いる．

口癖が頻繁に発生している音声では，視聴している

†名古屋工業大学大学院情報工学専攻

[1]https://www.jmooc.jp/

[2]https://www.khanacademy.org/

生徒は口癖が気になり集中力を欠いてしまう恐れがある．本研究では，口癖として「えーと」や「あのー」などの有声休止を検出し削除を行う．検出において有声休止の音素で構成される他の単語も存在するため音声認識のみの判定では誤検出を行う可能性が高い．そこで，本研究では，声道特性の特徴量から他の単語との区別を行い，有声休止の声道特性から判定を行う．また，ユーザによって有声休止の特徴が異なるため，あらかじめユーザには削除したい有声休止を見本選択してもらい，類似した有声休止を検出する．

本システムでは，映像の変化点の有無に基づき，音声の編集方法を制御することで，映像と音声を同期させている．映像の変化点検出は，類似画像判定に基づいており 3.5 節で詳述する．映像の変化点において，音声の編集を行う場合は，適宜，音声を伸長・削除する．

3. 講義映像の半自動編集

3.1 音声区間の取得

音声処理の性能を大きく左右するため記録した音声に対して，音声区間の検出を行う．本研究では，大語彙連続音声認識エンジンである Julius[3] を用いて編集する音声に対して音声区間の検出を行う．Julius では，音声信号の振幅と零交差数に基づいて音声の開始位置と終了位置を検出する．さらに，振幅と零交差によって検出された音声情報に対してガウス混合分布モデル (GMM : Gaussian mixture model) を用いて音声区間の判定を行う．音声 GMM と非音声 GMM から，短時間フレームごとに特徴量を抽出する．各 GMM の尤度計算を行い，音声 GMM と非音声 GMM の尤度比から音声区間の開始・終了を判定し，音声区間を取得する．

3.2 言葉の間の自動調整

言葉の間の自動調整を行うために音声信号に対して音素情報の取得を行う．Julius の単語・音素セグメンテーションキット[4]を用いて音声情報に対して音素のラベリングを行う．ラベリングされた音声から音素情報を取得する．Julius で得られる音素情報から講義映像の発話内容の品詞を解析し，接続詞の後，名詞の列挙による体言止めおよび名詞＋助詞の後に言葉の間が取られているか判定を行う．さらに音素情報から音素の時間区間を算出し，言葉の間が 0.5 秒取られているかを判定し，0.5 秒未満の場合は，0.5 秒になるまで無音を追加する．さらに，言葉の間が取るべきところではない言葉の間がある場合は，言葉の間の削除を行う．

3.3 早口の自動調整

本手法では，講義映像の音声区間情報および音素情報から 1 秒当たりのモーラを算出し，基準のモーラに対して発話速度の変更割合を算出する．i 番目の音声区間において，発話速度の変更割合の R_i は，t_i を音声区間時間，p_i を音素数，m をモーラ基準数として式 (1) で示す．

$$R_i = \frac{t_i}{p_i} m \tag{1}$$

本研究では，5 mora/s を聞き取りやすい基準として定め $m = 5$ として早口の自動調整を行う．早口の調整手法として発話速度の変更割合の R_i を基準にピッチ同期重畳加算 (PSOLA : Pitch Synchronous Overlap and Add)[2] を用いて，音響信号の声高を変えずに音の伸縮を行う．PSOLA は，波形の零交差点であるピッチマーク毎にハミング窓で波形を切り出し，ピッチ周期とフレーム時間長に合わせて切り出した波形を重畳加算し，合成波形を生成する．

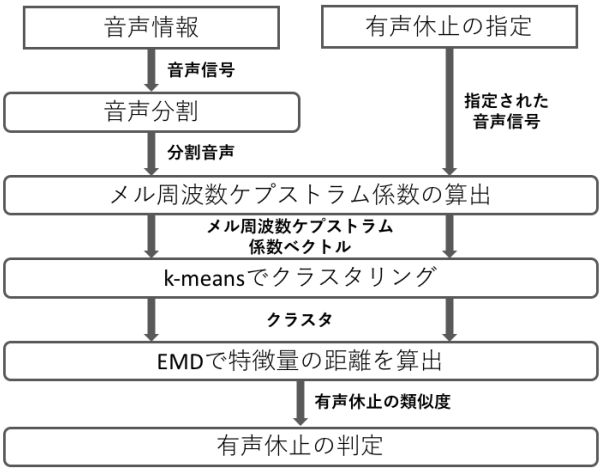

図 1: 有声休止の検出

3.4 有声休止の検出

有声休止の検出の処理の流れを図 1 に示す．本手法では，声道特性の特徴量から有声休止の検出を行うためメル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficients)[3] を求める．まず，ユーザが指定した見本の有声休止と音声情報を分割した分割音声からメル周波数ケプストラム係数ベクトルを算出する．次に，算出されたメル周波数ケプストラム係数ベクトルが 20 次元のベクトルがフレーム数分あり，膨大なベクトル数になるため，k-means でベクトルのクラスタリングを行う．見本の有声休止の音声 P およびクラスタの数を m とすると式 (2) で示される．

$$P = \{(\mu_{p_1}, \sum_{p_1}, w_{p_1}), \cdots, (\mu_{p_m}, \sum_{p_m}, w_{p_m})\} \tag{2}$$

最後に，クラスタリングされた特徴量の距離尺度を EMD(Earth Mover's Distance)[4] で算出し有声休止と類似しているか判定する．見本の有声休止の音声 P と分割音声 Q のクラスタの距離を $d_{p_i q_j}$，W を最小にする $f_{p_i q_j}$ とし距離尺度を EMD を式 (3)(4)(5) で示す．

[3]http://julius.sourceforge.jp
[4]http://julius.osdn.jp/index.php?q=ouyoukit.html

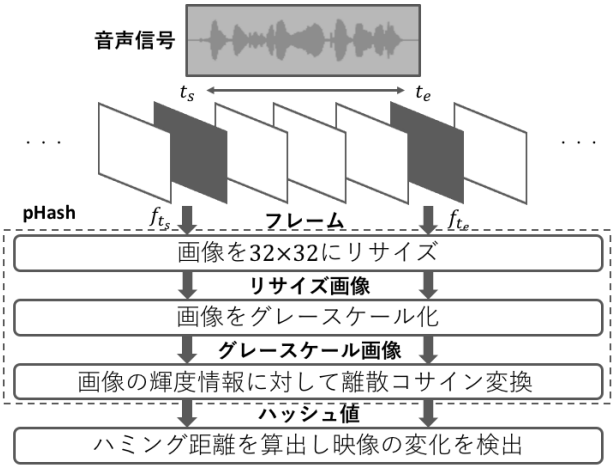

図 2: 映像の変化検出

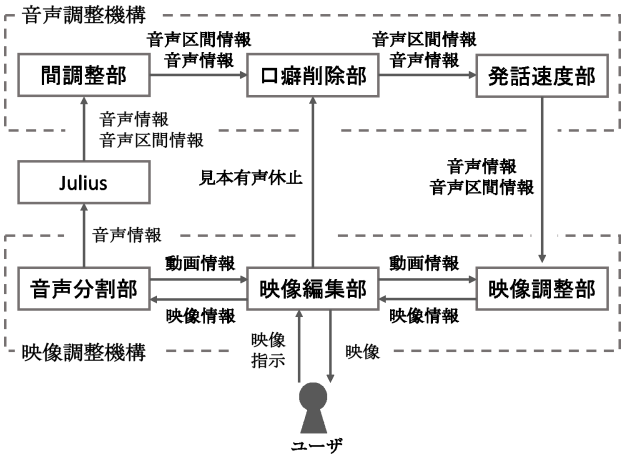

図 3: システム構成図

$$d_{p_i q_j} = \frac{\sum_{p_i}}{\sum_{q_j}} + \frac{\sum_{q_j}}{\sum_{p_i}} + (\mu_{p_i} - \mu_{q_j})^2 (\frac{1}{\sum_{p_i}} + \frac{1}{\sum_{q_j}}) \quad (3)$$

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j} \quad (4)$$

$$EMD(P,Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{p_i q_j}} \quad (5)$$

EMD が小さいほど見本の有声休止と分割音声が類似している．分割音声において特徴量の距離尺度から有声休止を検出し削除を行う．

3.5 映像の自動調整

映像の自動調整について図 2 を用いて述べる．編集される音声情報の音声区間および音素区間の時間の始端 t_s および終端 t_e とし，映像フレーム f_{t_s} および f_{t_e} を抽出する．映像の変化は，f_{t_s} および f_{t_e} に対して類似画像判定 Perceptual Hash(pHash)[5] を用いて検出を行う．pHash により映像フレームの 64 ビットのハッシュ値である h_{t_s} および h_{t_e} を算出する．映像の変化を求めるため，式 (6) で示すように h_{t_s} および h_{t_e} からハミング距離 d を求める．

$$d = \sum_{i=1}^{64} (h_{t_s i} + h_{t_e i}) mod(2) \quad (6)$$

$d \neq 0$ のとき，映像に変化があると判定され，音声の引き延ばしが発生した場合，始端 t_s および終端 t_e に合わせ映像の引き延ばしを行うことで整合性を保つ．音声の削除が発生した場合は，映像の削除を行わず音声と映像の整合性を保つ．

4. 講義映像作成支援システム

本節では，音声の早口，言葉の間および口癖を半自動的に調整を行う講義映像作成支援システムについて述べる．本システムは，反転授業作成のために撮影した講義映像に対して音声明瞭性を考慮した映像に半自動的に編集を行う．図 3 では，本システムのシステム構成を示す．本システムは，映像調整機構と音声調整機構で構成されている．

映像調整機構は，映像編集部，映像分割部および映像調整部で構成されている．映像編集部では，ユーザが映像を入力し見本有声休止などの編集指示を入力する．映像分割部では，入力された映像から動画情報と音声情報を分割し，音声情報を Julius に渡す．Julius では，音声区間情報および音素情報を出力する．映像調整部では，動画情報と自動調整された音声情報で映像を出力する．また，動画情報と音声区間情報および音素情報から映像の変化を算出する．映像の変化から動画情報が調整され，最後に音声情報と動画情報が合成された映像が出力される．

音声調整機構は，発話速度部，口癖削除部および間調整部で構成されている．発話速度部では，音声区間情報および音素情報から速度変更割合を算出し，PSOLA により音声情報の発話速度を調整し口癖削除部に渡す．口癖削除部では，ユーザから指定された有声休止を見本有声休止として全体の音声情報から類似している音声を検出し，削除を行う．間調整部は，音声区間情報および音素区間情報から言葉の間が必要な箇所を検出し，発話の間が 0.4 秒未満の短い間の場合，0.5 秒に自動調整を行う．

5. 講義映像作成支援システムの評価と考察

本節では，講義映像作成支援システムの評価と考察について述べる．本システムの実行例を示し，口癖削除についての予備実験を行い，実行例と予備実験から本システムを考察する．

5.1 実行例

図 4 を用いて本システムの実行例について述べる．本実行例は，スライド形式の講義映像の編集を行ってお

図 4: 実行例

り，図 4 の (A) では，講義映像が表示されている．図 4 の (B) では，本システムに入力した映像の元の音声情報の波形が表示されている．ユーザは，元の音声情報から見本の有性休止の範囲を選択することができる．見本の有性休止の選択後，全体の音声情報から類似している音声を検出し削除を行う．本実行例では，図 4 の (B) において有声休止を選択しており，図 4 の (C) では，有声休止が削除された音声情報が表示されている．また，早口である音声区間に対して引き延ばし，言葉の間が詰まっている部分を 0.5 秒まで言葉の間を広げた．図 4 の (D) では，修正された音声情報に合わせて動画情報が修正され，講義映像のサムネイル画像が表示されている．

5.2 予備実験

A 「本節では，(えーと) 講義映像作成支援システムについて (えーと) 述べる．」

B 「本システムは，(えーと) 反転授業の作成のために (えーと) 撮影した．」

C 「映像に対して (えーと) 音声明瞭性を考慮した (えーと) 半自動的に編集する．」

本システムの口癖検出の予備実験について述べる．本実験では，3 つの 10 秒間の音声，B および C を用意し，それぞれに「えーと」の有声休止を 2 つずつ含ませている．まず，A の音声を本システムに入力し，片方の有声休止を見本の有声休止として選択し有声休止の削除を行なったところもう片方の有声休止は削除することができた．しかし，B および C の音声では片方の有声休止を選択しても，もう一方の有声休止を検出し削除することができなかった．ここで，3 つの音声を一つの音声とし，A で選択した見本の有声休止で B および C の有声休止を検出したところどちらの有声休止も検出し削除することができた．

5.3 考察

本実行例および予備実験から本システムの考察を行う．本実行例より，本システムを用いることで講義映像を半自動編集することが可能である．早口調整および言葉の間調整では，声高を変えずに発話速度を変えることができ，聞き取りやすい音声となった．また，口癖削除では，本実行例よりユーザが削除したい有声休止を選択することで他の有声休止を削除することができた．予備実験より，指定した有声休止によって，有声休止の検出精度に大きく影響することが考えられる．そこで，今後，過去に検出したユーザの有声休止から機械学習を行い，個人に特化した有声休止の見本を形成し精度の向上および自動化を行う．

6. おわりに

本稿では，反転授業における講義映像の作成および編集を行う講義映像作成支援システムを提案した．本システムは，撮影された映像の言葉の間，音声の早口および口癖を検出し調整を行い生徒が聞き取りやすい音声明瞭性の高い音声に半自動的に編集を行う．本システムを用いることにより教師が円滑に講義映像を作成することができた．今後，ユーザの有声休止を機械学習を行い，個人に特化した有声休止の見本を形成し，有声休止の検出の精度を向上させる．

参考文献

[1] 横井聖宏, 馬場康輔, 須藤秀紹, 山路奈保子. 発話中の「間」がプレゼンテーションに対する聴衆の支持に与える影響 書評ゲーム『ビブリオバトル』の発表音声録音データ分析による考察. 日本感性工学会論文誌, Vol.15, No.3, pp.363-368, 2016.

[2] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Speech Communication 9, pp.453-467*, 1990.

[3] Md.Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. In *Speech Communication 54, pp.543-565*, 2012.

[4] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. In *Intl Journal of Computer Vision, Vol.40, pp.99-121*, 2000.

[5] Vipul Bajaja, Sanket Keluskar, Ravi Jaisawala, and Prof. Rupali Sawantb. Plagiarism detection of images. In *International Journal of Innovative and Emerging Research in Engineering, Vol.2, pp.140-144*, 2015.