6D-07

筆記音と手書き板書動画の同期による講義ビデオの音ズレ修正

周　東誠　　松崎　拓也
東京理科大学　理学部第一部　応用数学科

1.はじめに

電子ペンを用いた板書と音声解説による講義を動画として収録する際に、音声と映像の一方が他方より遅延する、いわゆる音ズレが頻発する。人手での修正はある程度まで可能だが、長時間の動画で遅延が徐々に蓄積していく場合などは、正確な修正は困難である。そこで、音ズレを自動的に修正したい。

本論文では、タブレット上での板書をメインとする講義動画を処理対象とする。電子ペンで字を書く際の音（筆記音）を、動画上でのペンストロークの開始時刻と同期させることで、音ズレが修正できると考えられる。長時間の動画における、筆記音とペンストロークの全てのアライメントを計算することは、原理的には可能だが、計算コストが非常に大きい。そこで、20秒ほどの区間におけるアライメント（局所アライメント）を、5〜10分おきに複数回計算し、最後にそれらの結果を元に、動画全体に対する音ズレの修正を行う。

2.多重HPSSによる筆記音の抽出

筆記音は打楽器音と類似しており、スペクトログラムにおいて、周波数方向に滑らかであり、時間軸方向に不規則な並びになっている。打楽器・コード楽器・歌声の音の特徴の違いを利用して、これらを分離する手法として、多重HPSS（Harmonic/Percussive Sound Separation）[1]がある。本研究では、特に実装が容易な直列多重HPSSを利用して筆記音の抽出を行った。

歌声ではなく話し声に対応するため、以下の工夫を行なった。筆記音は主として打楽器音（P成分）に分類される。人の話し声は、STFTの窓サイズを小さくし、周波数方向での特徴を抽出しやすくする事で、P成分から分離できると考えられる。橘ら[1]は、多重HPSSによって歌声を抽出する際、１回目のHPSSに用いるSTFTの窓サイズを200ミリ秒に設定し、抽出されたP成分に対する２回目のHPSSでは、窓サイズを30ミリ秒に設定することで、定常音成分（H成分）として歌声を得ている。一方、本研究では1回目のHPSSに用いる窓サイズを30ミリ秒に設定し、P成分として抽出された筆記音と多少の話し声の混合を、２回目により短い3ミリ秒の窓サイズで処理することで、話し声の成分を大きく抑制した。最後に得られたP成分の波形を、局所アライメントでの最大値で割って正規化し、閾値1/3を超える時点を筆記音の立ち上がり時刻として検出した。

3.フレーム間差分によるストローク開始時刻の検出

ペンのストロークが開始した時刻を検出するために、連続する２フレームの差分を用いた。単に差分画像の輝度の合計が一定値を超える点を検出すると、画面のスクロールや投げなわツールによる移動が検出されることもあるが、これらはペンストロークに比べて遥かに大きな差分を生ずる。このため、αとβを定数として、前フレームとの差分がα以上β以下となるフレームを、ストロークが行われている時刻として検出した。適切にαとβを設定する事で、１つのストロークが継続している間、フレーム差分は常にα以上、β以下となる。この区間の先頭を抽出する事で、ストローク開始時刻を検出した。

4. 筆記音とストロークのアライメント

動画全体の各時刻における音声・画像のズレ時間を計算したい。これを近似的に実現するために、一定時間 n 分おきに、20秒間の音声・画像を取り出し、そこから抽出した筆記音の立ち上がり時刻とストローク開始時刻の間の最適なアライメントを計算する。この局所的なアライメントを計算する際は、ズレの蓄積は考慮せずに、20秒間の音声・動画を同期させる最適なズラし時間を求める。前述のように、動画全体に渡ってズレは蓄積していくため、さらに先の局所アライメントでは、新たなズラし時間を求める。これを繰り返す事で、動画全体のおおよそのアライメントが得られる。以下、局所アライメントの手続きの詳細を述べる。

Audio-Video Out-of-Synch Correction for Lecture Videos by Aligning Pen Sound and Motion
Tosei SHU and Takuya MATSUZAKI・Tokyo University of Science

筆記音の立ち上がり時刻を$\{s_1, s_2, \ldots, s_m\}$、ストローク開始時刻を$\{f_1, f_2, \ldots, f_n\}$とする。基本的な考えは、各$f_i$に対して相手となる$s_j$を探す事である。具体的には、$\{f_1, f_2, \ldots, f_n\}$を後ろに一定時間$b$秒だけズラしたときに、各$f_i + b$に対し、それと最も近い$s_j$が、元々のストローク$f_i$に対応する筆記音の発生時刻だと考える。$f_i + b$と直近の$s_j$との差を$e_i(b) = \min_{1 \le j \le m} |f_i + b - s_j|$と書き、その和の$E(b) = \sum_{i=1}^{n} e_i(b)$をズレ時間$b$に対する損失として、これを最小にする$b$を求める。$e_i(b)$のグラフはのこぎり型のグラフになり（図１）、$e_i(b)$の和である$E(b)$は再び区分線形であるため、最小値を与えるズラし時間bを容易に求める事ができる。

5. 実験設定

実験の対象とした講義動画では、全て音が映像より遅れていたため、アライメントの計算では音が遅れていること（$b > 0$）を仮定した。また、より正確に修正するために、収録開始時に「川」などストロークが検出されやすい筆記を３、４回行った。入力とした講義動画は、iPadの画面収録機能を用いて作成した。表１に実験で用いた講義動画３本の基本情報を示す。動画は、軽量化ソフトHandBreakで予め容量を小さくしてから実験を行った（約800MBから約200MBへ）。局所アライメントを行う間隔は、$n = 10, 15$分の２パターンを比較した。

6. 実験結果

図２に、局所アライメントによって得られた最適なズレ時間の推移を示す。青い点は、人手により動画のズレ時間を検出したものであり、青い線はそれらに対する最小２乗法による回帰直線である。

図から、ズレは10分で0.3秒程度蓄積される事がわかる。しかも、必ずしも線形に蓄積していくのではなく、グラフ(B)のように非線型となることもある。各局所アライメントで得られたズレ時間と人手による修正結果の回帰直線との絶対誤差／２乗誤差の平均の値を表２に示す。概ね人手によるものと近い結果を得る事ができたが、0.1秒程度の誤差を生じる部分も見られる。

動画(A)と(B)に対する結果で、極端に外れた値が見られる。原因として、HPSSで抽出されるP成分の特性を決めるパラメータω_Pの影響が考えられる。動画(A)に対し、ω_Pを変化させたときの局所アライメントの結果を表３に示す。改良法として、ω_Pの動的な調整や、直列多重HPSSとは異なる方式による多重HPSSの適用が考えられる。

表１：実験に用いた動画

	動画の長さ	先頭のズレ時間（秒）
(A)	1:30:11	5.697
(B)	1:18:41	1.429
(C)	58:43	27.584

表３：ω_Pの値による動画（A）の局所アライメント結果の変化

動画(A)	$\omega_P = 0.5$	$\omega_P = 0.7$	$\omega_P = 1.0$	人手
10分	0.01	0.30	0.78	0.30
20分	0.00	0.59	0.62	0.60
30分	0.00	0.87	0.87	0.89

表２：修正の結果

	n	絶対誤差平均	2乗誤差平均	n	絶対誤差平均	2乗誤差平均
(A)	10	0.151	0.032	15	0.118	0.022
(B)		0.057	0.004		0.061	0.008
(C)		0.062	0.005		0.067	0.006

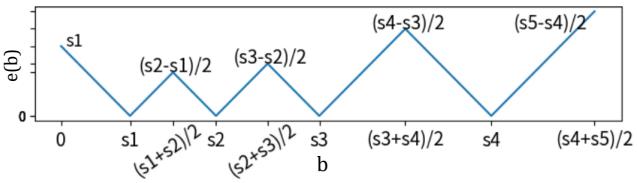

図１：$e_i(b)$のグラフ

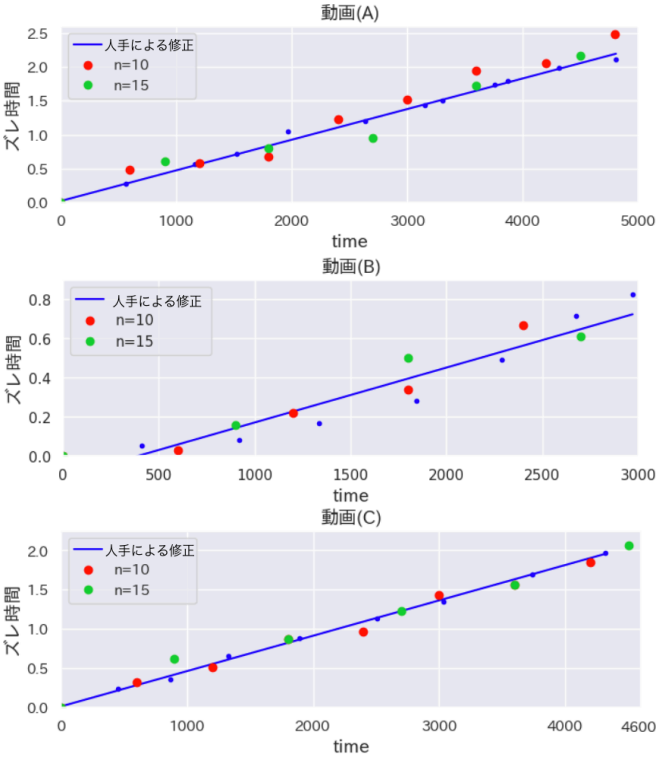

図２：自動修正の結果

参考文献

[1] 橘 秀幸，小野 順貴，嵯峨山 茂樹．スペクトルの時間変化に基づく音楽音響信号からの歌声成分の強調と抑圧．情報処理学会研究報告，Vol.2009-MUS-81，No.12，2009 / 7 / 30.