2R-06

音声認識を用いた講義音声の早口分析

池川 心 †
和歌山大学システム工学部 †

西村 竜一 ‡
和歌山大学データ・インテリジェンス教育研究部門 ‡

1 はじめに

大学の講義・議論・学会発表など，話者の説明のスピードが速い（早口）と，聴講者は聞き取りにくく，言いたいことが聴講者には正しく伝わらないと指摘されることが多い．一方，人間の心理的状況は無意識に発話速度に影響する．興奮して早口になっていても，話者自身がそれに気が付かないことも多い．

先行研究 [1] では，Web Speech API[*1]を用いて発話を書き起こし，1 秒あたりの発話文字数を計測し，早口を検出した．しかし，早口と検出した部分と聴講者が早口と感じたタイミングに異なりがあった．

2 実験概要と使用データ

本研究では，次に概要を示す実験を行った．早口を可視化して，話者の早口の改善につなげるための手法を検討することが目的である．

- 【実験 1】自動音声認識を用いた発話速度の計測．音声認識の精度を補うために，複数の音声認識 API の利用を試みた．
- 【実験 2】早口を表現する発話速度以外の物理特徴の調査．a) 声量変動，b) 音高変動を音声信号から抽出して調査した．
- 【実験 3】話者の心理的状況が早口に与える影響の調査．アンケートを通じて聴講者（大学生）が感じた印象の主観的評価を行った．

対象としたのは，和歌山大学のオンデマンド型オンライン授業の講義音声（録音）である．話者数（教員数）は 6（A〜F），講義数（音声ファイル数）は 16（A-1〜F-16）である．なお，本研究では，人手で判断して，1 つの講義から前半・中盤・後半の音声を切り出し，それぞれを 12〜46 秒とした．切り出した音声ファイルを 2 秒間の無音を挟んで連結し，各講義を 59〜135 秒の音声ファイル 1 つにまとめている．

アンケートによって，授業の聴講者を想定した大学生（実験協力者 16 名）が，前述の講義音声を聴取して印象として感じた早口の程度の定量化を試みた．協力者に提示した設問を表 1 に示す．設問は，5 つの大項目（合計設問数 16）で構成した．大項目 1〜4 の回答は，「5. 非常に同意できる」〜「1. 全く同意できない」の 5 段階（リッカート尺度）とした．大項目 5 の回答は，文章による自由記述とした．

なお，設問の中には，【実験 3】で使用する話者が抱く 6 つの心理的状況（「楽しみ（興奮，喜び）」「嫌気（嫌悪，苦手）」「悲しみ（悲哀，無力）」「恐れ（パニック，不安，緊張感）」「怒り（激昂，苛立ち）」「焦り」）に関する項目を含んでいる．

3 【実験 1】自動音声認識を用いた発話速度の計測

【実験 1】では，音声認識 API を用いて講義音声の自動書き起こしを行った後，ひらがな化 API[*2]でひらがな変換をした．ひらがなの文字数を数え上げて，1 秒あたりの発話文字数を計測した．本研究では，この結果の値を発話速度（[文字数/秒]）とする．

表 2 に発話速度の結果を示す．Google Speech Recognition API と Google Cloud Speech-to-Text API[*3]の 2 つの API を試し，それぞれの結果を使用したときの発話速度を掲載する．表中の「手動」は，比較のため，人手で書き起こしたときの結果である．括弧内の数字は，各 API の出力の「手動」に対する文字数（ひらがな）の割合（[%]）である．

Google Cloud Speech-to-Text API を用いた場合，Google Speech Recognition API に比べて，若干の高い精度を確認したが，専門用語や間投詞等の誤認識

Analysis of fast speaking in lecture speech based on speech recognition

† Kokoro Ikegawa, School of Systems Engineering, Wakayama University
‡ Ryuichi Nisimura, Data Intelligence Education Research Division, Wakayama University

[*1] https://developer.mozilla.org/ja/docs/Web/API/Web_Speech_API

[*2] https://labs.goo.ne.jp/api/jp/hiragana-translation/

[*3] https://cloud.google.com/speech-to-text?hl=ja

表1　アンケートで協力者に提示した設問

項目1　聞き取りやすさ・理解のしやすさについて
設問1-1　話者の声は, 聞き取りやすい.
設問1-2　話者の説明は, 理解しやすい.
項目2　発話速度について
設問2-1　話者の話す速さは, 速い.
設問2-2　話者の話す速さは, 適切である.
設問2-3　話者の話す速さは, 遅い.
項目3　話し方の特徴について
設問3-1　話者の声の大きさの変化は, 大きい.
設問3-2　話者の声の高さの変化は, 大きい.
設問3-3　話者は, 淡々と話している.
設問3-4　話者は, 一生懸命話そうとしている.
項目4　話者の心理的状況について
設問4-1　話者は, 楽しみ(興奮, 喜び)を感じている.
設問4-2　話者は, 嫌気(嫌悪, 苦手)を感じている.
設問4-3　話者は, 悲しみ(悲哀, 無力)を感じている.
設問4-4　話者は,恐れ(パニック, 不安, 緊張感)を感じている.
設問4-5　話者は, 怒り(激昂, 苛立ち)を感じている.
設問4-6　話者は, 焦りを感じている.
項目5　その他
上記以外に, 音声を聞いて受けた印象がありましたらご記入ください. (記述)

が生じた．その結果，正確に書き起こした場合よりも出力される文字数（ひらがな）が少なくなった．人手と比較して，全体平均で減少が 20% となった．特定の講義（D-6, C-7, F-15, F-16）では，自動音声認識が中断し，文字数が著しく減少していた．この講義音声を確認したところ，発話内の声量（声の大きさ）の変動が大きかった．よって，この減少率は，早口の可視化のファクタとして利用できる可能性があると考える．ただし，基準となる正確な（人手相当の）書き起こしを用意することができないため，2 つの API 出力の比較等を詳細に検討する（2 つの出力の相対的な値を尺度とする）必要がある．

4　【実験 2】早口を表現する発話速度以外の物理量の調査

【実験 2】では，物理的特徴として声量と音高を調査した．音声信号を切り出したフレーム内の振幅値（絶対値）の平均について，前フレームとの差分を求め，さらに 1 つの音声ファイルを通じて，その差分値を平均したものを「a) 声量変動」とした．また，音声信号から WORLD[2](D4C edition[3]) の DIO を用いて基本周波数（F0）を推定し，その値を正規化した．算出した値のフレーム間の差分を求めて平均した値を「b) 音高変動」とした．音声信号のサンプリング周波数は 44.1 kHz，フレーム長は 128 である．

アンケートを通じて定量化した早口の程度との相関を求めた．a) 声量変動と早口の程度の相関係数は-0.62 であり，声量変動が小さい場合，早口に聞こえることがわかった．b) 音高変動と早口の程度との相関は確認できなかった（相関係数：-0.12）．

表2　音声認識 API と手動書き起こしの比較, 発話速度 [文字数/秒] 及び手動に対する音声認識 API 出力の文字数の割合（括弧内）[%]

ファイル名（話者-講義）	手動	API平均	Google Speech Recognition API	Google Cloud Speech-to-Text API
A-1	9.04	8.00 (82.6)	7.52 (83.1)	7.43 (82.1)
A-2	8.58	7.84 (87.1)	7.49 (87.3)	7.45 (86.8)
A-3	9.00	7.95 (82.4)	7.51 (83.4)	7.34 (81.4)
B-4	6.99	6.41 (87.5)	6.16 (88.1)	6.08 (87.0)
B-5	6.76	6.34 (91.0)	6.15 (91.2)	6.12 (90.8)
C-6	6.51	5.44 (75.2)	5.14 (79.0)	4.66 (71.4)
C-7	6.82	5.58 (72.4)	4.61 (67.5)	5.29 (77.3)
C-8	6.76	5.56 (82.1)	5.47 (80.9)	5.63 (83.2)
D-9	6.59	6.29 (93.6)	6.15 (93.7)	6.14 (93.4)
D-10	5.73	5.28 (88.4)	4.98 (87.0)	5.13 (89.7)
D-11	6.28	6.17 (97.2)	6.12 (97.4)	6.10 (97.0)
E-12	6.18	5.93 (93.8)	5.85 (94.5)	5.76 (93.1)
E-13	5.87	5.72 (95.9)	5.63 (95.8)	5.65 (95.9)
E-14	6.87	6.55 (92.9)	6.41 (93.2)	6.37 (92.6)
F-15	7.10	5.17 (58.5)	3.86 (53.7)	4.54 (63.3)
F-16	7.81	4.71 (41.8)	2.57 (34.7)	3.74 (48.8)
平均	**7.06**	**6.21 (82.6)**	**5.73 (81.9)**	**5.84 (83.4)**

5　【実験 3】話者の心理的状況が早口に与える影響の調査

【実験 3】では，話者が「焦り」を感じていると早口になる傾向を確認した．早口と「焦り」の程度（聴講者の印象による）の相関係数は-0.82 であった．

6　今後の課題

協力者（大学生）による主観評価において，話者（教員）を特定できてしまうことによる先入観の影響を調査する必要がある．協力者（大学生）がアンケートを通じて回答した印象が，日々接する話者（教員）のパーソナルティに影響を受けている恐れがある．音声匿名加工処理の適用や話者（教員）を知らない協力者による評価等によって，先入観を排除して比較する必要があると考える．また，講義の内容や難易度の影響も生じている可能性が残っている．

謝辞 本研究は，JSPS 科研費 JP18K02862, JP21K12155 の助成を受けたものです．

参考文献

[1] 松浦ら, 情報処理学会全国大会論文集, vol.2017, no.1, pp.913–914, 2017.
[2] M. Morise, et.al, IEICE trans. on information and systems, vol.E99–D, no.7, pp.1877–1884, 2016.
[3] M. Morise, Speech Comm., vol.84, pp.57–65, 2016.