Vol.2016-MUS-111 No.30
2016/5/21

音声の韻律分析及び表情の特徴抽出による面接支援システム

棚橋 徹[1,a)]　高屋敷 弓恵[1,b)]　北原 鉄朗[1,c)]

概要：就職活動において重要な要素の 1 つに面接が存在する．面接では, 視線, 表情, 抑揚の有無, 話し方などで就職志願者を精査している．そのため, 面接の練習はとても重要である．しかし, 1 人で練習するのは難しい．そこで我々は, 音声の韻律分析及び表情の特徴抽出による面接支援システムを提案する．このシステムでは, あらかじめ用意された文章を PC の前で読み, その様子を録音, 録画する．それに対して音声, 画像処理を用いて解析, 見本となるデータとの比較を行い, ユーザーに視線, 笑顔, 顔の動き, 話速度, 全体の抑揚, 強調の 6 項目についてフィードバックをする．本システムを使用することにより, 面接で自分が強調したい部分が強調できるようになることを期待する．実験では, システムの評価と人による評価に差異がないかを確かめるためにアンケートを行った．その結果, 抑揚と強調に関しては相関係数がそれぞれ 0.68 と 0.42 であった．ただし, その他の評価項目では相関係数が 0.2 を下回った．

キーワード：面接, 支援システム, 韻律分析, 表情

1. はじめに

面接において重要とされるのは, 表情や仕草, 視線, 話し方, 質問への答え方などが挙げられる [1], [2]．しかし, 1 人で練習しても面接官にどのように見られているかわからない人や, 人に見てもらうと恥ずかしさや煩わしさを感じる人は多く存在する．そのため, 面接の練習があまりできない人は多いと考える．

1 人で面接やプレゼンテーションの練習を行える既存システムの 1 つに「プレゼン先生」[3] がある．プレゼンテーション練習時の様子を録音・撮影し, リアルタイムに基本周波数 (以後 F0) による抑揚や顔の向きなどを取得し, それぞれ閾値を越えた場合ユーザーに警告がされる．更に, 発表後にグラフの表示などで過去との比較も可能である．また,「VR 面接：坂本龍馬と面接」[4] は, 視線をヘッドマウントディスプレイで, 音声をヘッドセットのマイクから取得し, 適切な職種を紹介してくれる．両システムも , 話す内容に依存せず, 全体的に抑揚があるのかなどのフィードバックを行う．他にも, 面接支援システムとして [5], [6], [7] が開発された．しかし, これらのシステムでは, 自分が強調したいところが強調されているのかは判定していない．強調を判定するシステムとして, プレゼンテーションをする際に単語が強調できているかを判定するシステム [8] が開発された．このシステムでは, ある一文を話したときに, 指定した単語が強調できているかを判定するシステムである．しかし, いずれのシステムも面接という環境下で自分が主張したい箇所が強調できているのか, 表情が柔らかかったかなどの判定は行っていない．

そこで, 本研究では, あらかじめ用意した文章を読んでもらい, 強調すべき箇所が強調されているかを判定する就職活動向け自己 PR 練習システムを開発する．本システムでは, あらかじめ用意していた文章を読んでいる間の視線, 笑顔, 顔の動き, 話速度, 全体の抑揚, 強調の 6 つの点を見本データと比較してフィードバックを行う．見本データとの比較を行うことにより, 明るく元気に話すタイプの人や, 静かだが説得力のある話し方をするタイプの人でも適切なフィードバックを行うことが可能となる．また, フィードバックの際には, 六角形のグラフでの表示, 文章での良かった点・悪かった点を表示する．本システムを使用することで, 面接での自己 PR が苦手, 不安と感じる人の話し方や表情が改善されることを期待する．

2. システム概要

本システムでは, ユーザーにある決められた文章を読んでもらい, その時のユーザーの様子を音声・画像として記録する．記録されたデータから視線, 笑顔, 顔の動き, 話速度, 全体の抑揚, 強調の 6 項目に注目し, あらかじめ用意した見本となるデータと比較しフィードバックを行う．

[1] 日本大学
Nihon University
[a)] tanahashi@kthrlab.jp
[b)] takayashiki@kthrlab.jp
[c)] kitahara@kthrlab.jp

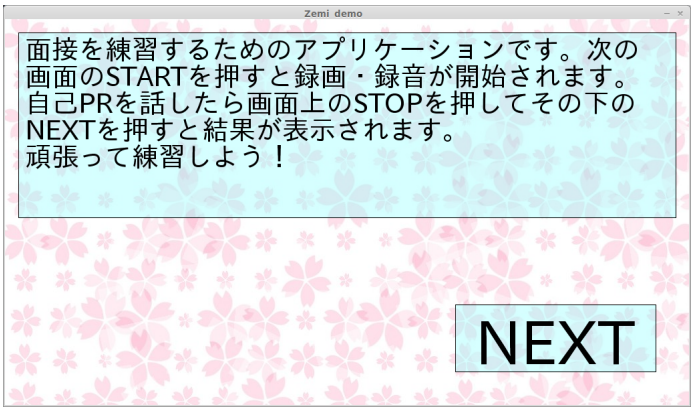

図 1 操作方法表示画面

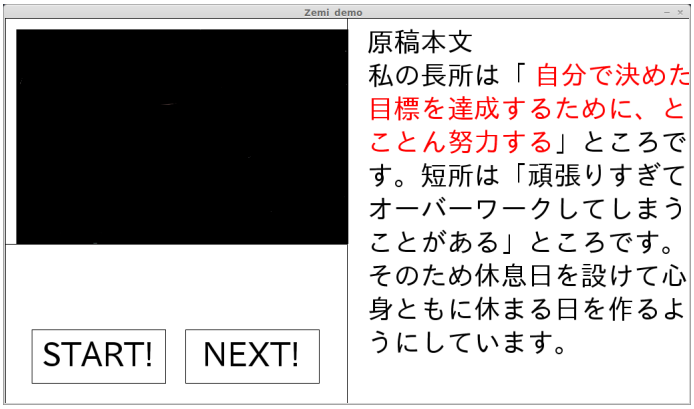

図 2 面接練習画面

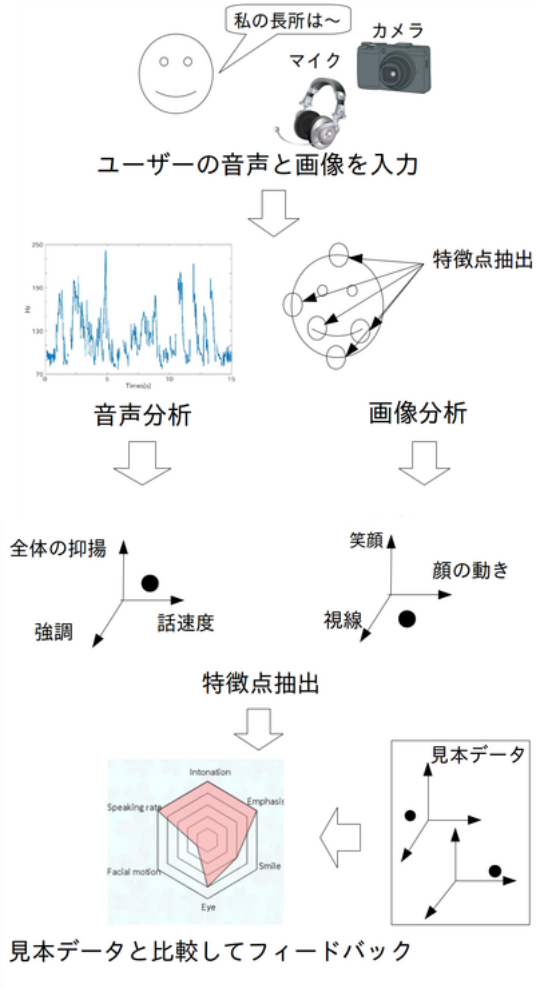

図 3 システムの流れ

2.1 システムの流れ

システムを起動すると, 操作方法が表示される (図 1). 操作方法を読み終えたら NEXT というボタンを押し, 面接練習に移行する (図 2). ここで, ユーザーには PC に対して正面に座ってもらい, PC 本体の web カメラにしっかりと自分が写っているかを確認してもらい, ヘッドセットを装着してもらう. そして, START を押してからユーザーに面接での自己 PR を想定してあらかじめ用意された文章を読んでもらう. 用意された文は以下である. 下線は強調を意識して読んでもらう部分である.

- 私の長所は「自分で決めた目標を達成するために, とことん努力する」ところです. 短所は「頑張りすぎてオーバーワークしてしまうことがある」ところです. そのため, 休息日を設けて心身ともに休まる日を作るようにしています.

このとき web カメラから画像を取得し, ヘッドセットのマイクから音声を取得する. 終了後, 画像処理・音声処理を行う. 処理の流れを図 3 に示す. ここでは、前述した 6 項目について解析, 評価し, フィードバックの表示を行う. フィードバックの画面では, 各項目について評価された値に応じてグラフ, 良かった点, 悪かった点が表示される.

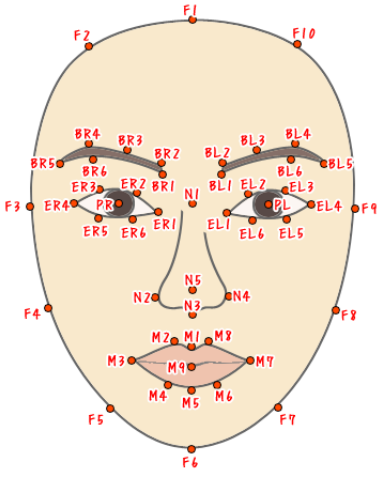

図 4 detectface(); で取得可能な特徴点 [9]

2.2 画像処理

Web API の detectface()[9] を利用し顔の特徴点 (**図 4**) の座標を取得する. 本システムでは PC の web カメラから入力された映像を 1 秒間に 1 回画像として記録し, その画像に対して detectface() を適用する. これを用いて視線, 笑

顔, 顔の位置検出を次の方法で行う.

- 視線は, 正面のカメラから視線が外れた割合を求める. 目端, 瞳の x 座標を検出しその差分を計算, また差分同士の比較をする. 左右の瞳が右, または左に偏っているかを次式により算出する.

$$\big||\mathrm{ER1}_x - \mathrm{PR}_x| - |\mathrm{PR}_x - \mathrm{ER4}_x|\big| > 2[\mathrm{px}] \quad (1)$$

かつ,

$$\big||\mathrm{EL4}_x - \mathrm{PL}_x| - |\mathrm{PL}_x - \mathrm{EL1}_x|\big| > 2[\mathrm{px}] \quad (2)$$

($\mathrm{ER1}_x$ は図 4 における ER1 の x 座標を表す. 他も同様である). これにより, 何回視線がカメラから外れたかを求め, 記録した画像の総数で割り, 視線が外れた割合を求める. この視線が外れた割合を $x_{\mathrm{u},1}$ とする.

- 笑顔は, 左右両方の口端が均等にカメラ起動時より上がっているかを計算する. すなわち,

$$|\mathrm{N2}_y - \mathrm{M3}_y| > |\mathrm{N2}_{y_0} - \mathrm{M3}_{y_0}| \quad (3)$$

$$|\mathrm{N4}_y - \mathrm{M7}_y| > |\mathrm{N4}_{y_0} - \mathrm{M7}_{y_0}| \quad (4)$$

$$\big||\mathrm{N2}_y - \mathrm{M3}_y| - |\mathrm{N4}_y - \mathrm{M7}_y|\big| < 10[\mathrm{px}] \quad (5)$$

($\mathrm{N2}_{y_0}$ はカメラ起動時の $\mathrm{N2}_y$ を表す. 他も同様) の時笑顔と判定し, 視線と同様に割合を求める. この笑顔でいた割合を $x_{\mathrm{u},2}$ とする.

- 顔の動きは, 顔の位置が動いた割合を求める. 顔の縦幅 FH_y, 顔の横幅 FW_x を,

$$\mathrm{FH_y} = \mathrm{F6}_y - \mathrm{F1}_y \quad (6)$$

$$\mathrm{FW_x} = \mathrm{F9}_x - \mathrm{F3}_x \quad (7)$$

とし, これを 1 時刻前に取得した画像との差分を求め動いたかを次式で判定する.

$$\mathrm{FH}_y - \mathrm{FH}'_y > 5[\mathrm{px}] \quad (8)$$

$$\mathrm{FW}_x - \mathrm{FW}'_x > 5[\mathrm{px}] \quad (9)$$

$$\mathrm{F3}_x - \mathrm{F3}'_x > 5[\mathrm{px}] \quad (10)$$

$$\mathrm{F1}_y - \mathrm{F1}'_y > 5[\mathrm{px}] \quad (11)$$

(FH'_y は 1 時刻前に取得した FH_y を表す. 他も同様.) これにより, 顔が動いた回数を求め, 視線と同様に割合を求める. この顔が動いた割合を $x_{\mathrm{u},3}$ とする.

2.3 音声処理

音声は話速度, 全体の抑揚, 強調 3 つについて評価を次のように行う.

- 話速度は, 録音した音声に対して, Julius[10] による強制アライメントを行い, その出力結果から, 発話した文全体の母音数と発話時間 (発話終了時刻 - 発話開始時刻) を取得し,

$$(\text{話速度}) = \frac{(\text{母音数})}{(\text{発話時間})} \quad (12)$$

により, 話速度を計算する. この話速度を $x_{\mathrm{u},4}$ とする.

- 全体の抑揚は, 録音した音声全体に対して, 5ms ごとに求めた振幅と F0 を用いる. F0 の推定には DIO[11] を用いる. F0 は cent 単位に変換する. その後, 振幅と F0 の時間変化の四分位幅を求め, 全体の抑揚とする. この時の振幅の四分位範囲を $x_{\mathrm{u},5}$, F0 の四分位範囲を $x_{\mathrm{u},6}$ とする.
- 強調は, 先ほど求めた振幅と F0 に対して, 強調区間に対応する区間とその直前の区間に対応する区間をそれぞれ抽出し両者の中央値の差を求める. 今回の文章の場合, 文頭の「私は」が強調区間のその直前の区間に相当する. この時の振幅の中央値の差を $x_{\mathrm{u},7}$, F0 の中央値の差を $x_{\mathrm{u},8}$ とする.

2.4 見本データとの比較

本システムでは, 文章を読む訓練を受けた複数人に対象文を読んでもらい, これを見本データとして用いることを想定している. 今, N 人の人に読んでもらったデータを $\{p_1, p_2, ..., p_N\}$ とすると, 各見本データ p_k に対して, 2.2, 2.3 節と同様の処理を行い, 特徴ベクトル $\boldsymbol{x}_{p_k} = (x_{p_k,1}, x_{p_k,2}, x_{p_k,3}, x_{p_k,4}, x_{p_k,5}, x_{p_k,6}, x_{p_k,7}, x_{p_k,8})$ を求め. これとユーザーに対する特徴ベクトル $\boldsymbol{x}_{\mathrm{u}} = (x_{\mathrm{u},1}, x_{\mathrm{u},2}, x_{\mathrm{u},3}, x_{\mathrm{u},4}, x_{\mathrm{u},5}, x_{\mathrm{u},6}, x_{\mathrm{u},7}, x_{\mathrm{u},8})$ をそれぞれ見本データとの差を求め, 0〜4 に離散化する. この際, この差が十分に小さいとき 0 とし, この差が十分に大きいとき 4 とする. 特徴量を 0〜4 に離散化する関数を $\mathrm{disc}_i()$ とすると,

$$\mathrm{E}_k = \sum_{i=1}^{8} \mathrm{disc}_i(|x_{\mathrm{u},i} - x_{p_k,i}|) \quad (13)$$

が最小となる k を求め, これを $\hat{k}$ とする. 特徴量の離散化するための閾値は, 特徴量の種類ごとに実験的に定めた.

2.5 結果の表示

結果の表示はグラフと文章の表示の両方を行う. 結果表示例を図 5 に示す. 2.4 節で求めた比較対象見本データ $p_{\hat{k}}$ との近さを 0〜4 の 5 段階で表示する. グラフ表示においては、見本データに近い (評価が高い) 方が高い値にした方が分かりやすいと思われるため, $4 - \mathrm{disc}_i(|x_{\mathrm{u},i} - x_{p_k,i}|$ を表示する. ただし, グラフ表示部が六角形なので, 全体の抑揚 $(x_{\mathrm{u},5}, x_{\mathrm{u},6})$ と強調の度合い $(x_{\mathrm{u},7}, x_{\mathrm{u},8})$ は F0 に関するものと振幅に関するものを 1 つにまとめて表示する. すなわち,

$$\left\lceil \frac{1}{2} \sum_{i=5}^{6} \{4 - \mathrm{disc}_i(|x_{\mathrm{u},i} - x_{p_{\hat{k}},i}|)\} \right\rceil \quad (14)$$

を表示する ($x_{\mathrm{u},7}, x_{\mathrm{u},8}$ も同様). ここで, $\lceil\ \rceil$ は天井関数を

表 1 文章によるフィードバック表示一覧

項目	評価値	結果表示
視線	3 ～ 4	しっかりカメラを向いて話すこと が出来ています．この調子で 他の項目も頑張りましょう．
	0 ～ 2	視線が左右に動いています． しっかりとカメラを見て 話すように心がけましょう．
笑顔	3 ～ 4	終始柔らかい表情で話すこと が出来ています．この調子で 他の項目も頑張りましょう．
	0 ～ 2	少し表情がぎこちないです． もう少し柔らかい表情で話すよう にしてみましょう．
顔が動き	3 ～ 4	落ち着いた様子で話すこと が出来ています．この調子で 他の項目も頑張りましょう．
	0 ～ 2	話しているとき顔がふらふらと 動いて落ち着きがないです． 背筋を伸ばして落ち着いた 様子で話すよう心がけましょう
話速度	3 ～ 4	適切な速度で話すことが出来ています． この調子で他の項目も頑張りましょう．
	0 ～ 2	話すスピードが早い (遅い) です． 落ち着いて話すようにしましょう．
抑揚	3 ～ 4	文全体で抑揚がつけられています． この調子で他の項目も頑張りましょう．
	0 ～ 2	棒読みになっていませんか？ もっと声をしっかりとだし， 句読点を意識しましょう．
強調	3 ～ 4	強調するべき場所が強調出来ています． この調子で他の項目も頑張りましょう．
	0 ～ 2	強調するべき場所が強調できていないです． 強調をつけるところを意識して 繰り返し練習してみましょう．

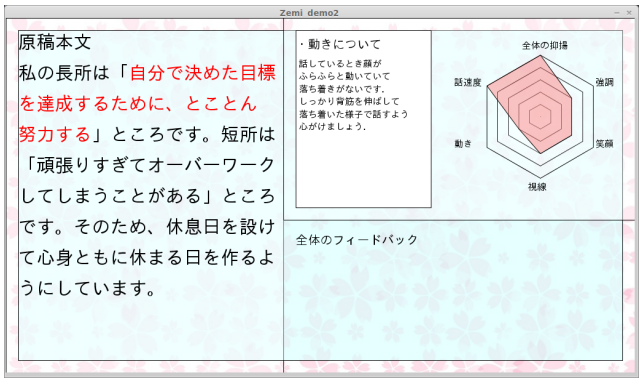

図 5 フィードバック画面

表す．文章の表示については，グラフの表示の際に求めた 5 段階評価を利用し，それに応じて文章が変化するようになっている．表示される文章については，5 段階評価した項目の中で一番高いものを表示する．表示される内容の一覧を表 1 に示す．

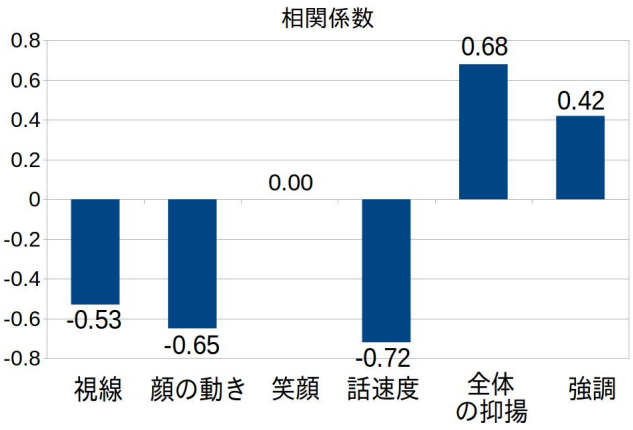

図 6 システム評価と人による評価の相関係数

3. 評価実験

評価実験では，本システムの各項目における 5 段階評価が適切かどうかを調べるために，システムの評価と人の評価の比較実験を行った．今回の実験では見本データに就職活動を経験した大学 4 年生 2 名 (男子 1 名，女子 1 名) を使用した．また，この 2 名はサークルや，ラジオなど話す経験が豊富であった．就職活動生を模した 5 人の被験者 (23~24 歳，男性 4 名，女性 1 名) に実際にシステムを使用してい，その様子をビデオカメラで正面から録画した．録画したものを被験者とは別の 8 人 (22~ 23 歳，男性 5 人，女性 3 人) に見てもらい，システムが評価する項目と同様の 6 項目について，5 段階評価 (0~4) をつけてもらった．その結果と，システムの評価した値に差異がないかを確かめ，相関があるかを調べた．相関が高ければ，システムの評価は人の評価と近いということが推察できる．

調べた結果を図 6 に示す．相関を調べた結果，全体の抑揚，強調では相関係数が 0.68 と 0.42 となり，相関があることがわかった．しかし，ある被験者に関して，システムでの評価全体の抑揚，強調共に 3 と良い評価が表示されているのに対して，人による評価では両項目とも 1 となっていた．これは，その被験者は，かすれたような声で人によっては聞きづらいため，その分人による評価が下がってしまったのではないかと考える．また，画像に関しては，正の相関がある項目は存在しなかった．視線は，被験者 5 人中 4 人に対するシステムによる評価が低く，人による評価は高かった．これは，見本となるデータとして使用した 2 名が，文章を読み間違えないように注力した結果，視線がカメラから外れている割合が高かったからではないかと考える．また，顔の動きに関しては，頷きなど説得力のある話し方をしていてもシステムでは顔が動いたと判定されてしまっていた．そのため，システムの評価は低いが，人による評価は高いということがあった．このように，動きの中でもプラスに働くものに関して，うまく判定できなかったと考えられる．

4. おわりに

本研究では, 就職活動未経験者を対象とした面接練習支援システムを提案した. 評価実験では, 全体の抑揚, 強調に関しては正の相関が見られたため, ユーザーに対して適切なフィードバックができるということがわかった. しかし, その他の項目に関しては, 正の相関が表れなかった. これは, 頷きなど説得力のあるように見える動きでも, システムでは動いたと判定されたからだと考えられる. また, 正の相関がでた全体の抑揚や, 強調の評価項目でも, システムによる評価と人による評価が分かれた被験者もいた. そのため, 見本データの増加や, 話し方のタイプの考慮について検討する必要があると考える.

今後の課題として 2 つあげられる. 1 つは取得している特徴についてである. 今回, 抑揚や強調といった部分は F0 と声の振幅で判定したが他にも間の使い方や前後の文章によって抑揚や強調をつけることが可能であると予想される. そのような要素を考慮して判定していく必要があると考える. もう 1 つは見本データについてである. 今回見本データには話す訓練を受けた大学生 2 名を使用したが, 面接において話し方のタイプは人それぞれあるのではないかと考えられる. 例えば, 元気のある話し方をする人もいれば, 慎重に言葉を選び, 説得力のある話し方もいることが考えられる. このような話し方のタイプについても今後考慮していく必要があると考える.

参考文献

[1] 坪田まり子, 就活必修！面接術 受け方・話し方・聞き方 2016, さくら舎, 2014.
[2] 渡辺 智美, 中村 亮太, 上林 憲行: 採用面接における非言語行動の印象改善方法の提案, 情報処理学会 第 75 回全国大会, 1ZE-8, 2013.
[3] 栗原 一貴, 後藤 真孝, 緒方 淳, 松坂 要佐, 五十嵐 健夫: プレゼン先生：音声情報処理と画像情報処理を用いたプレゼンテーションのトレーニングシステム, WISS 第 14 回 インタラクティブシステムとソフトウェアに関するワークショップ, pp. 59–64, 2006.
[4] VR 面接ー坂本龍馬と面接ー, 株式会社 DODA, URL:http://doda.jp/promo/campaign/mirainomensetsu/ (2015 年アクセス)
[5] T. Barur, T. Ionut, G. Patrick, P. Kaska, and A. Elisabeth.: A Job Interview Simulation: Social Cue-Based Interaction with A Virtual Character, *IEEE International Conference on Social Computing (SocialCom2013)*, pp. 220–227, 2013.
[6] J. Matthew, B. Laura, F. Micael, J. Neil, A. Michael, J. Emily, W. Katherine, O. Dale, and Morris, D, B.: Virtual Reality Job Interview Training for Veterans with Posttraumatic Stress Disorder, *Journal of Vocational Rehabilitation* 42, pp. 271–279, 2015.
[7] H. Tanaka, S. Sakti, Graham. N, T. Toda, H. Negoro, H. Iwasawa, and S. Nakamura: Automated Social Skills Trainer, *IUI '15 Proceedings of the 20th International Conference on Intelligent User Interfaces*, pp. 17–27, 2015.
[8] 小島淳嗣, 伊藤克亘, 花泉 弘: F0 最大値とアクセント成分最大値を用いたプレゼンテーション音声の重要性強調判定, 6Q-06, 2016.
[9] 顔検出ライブラリ detectface();, インクリメント株式会社,URL:http://www.increment.co.jp/product/detectFace/index.html (2015 年アクセス)
[10] 汎用大語彙連続音声認識エンジン Julius, URL：http://julius.osdn.jp/ (2015 年アクセス)
[11] 森勢将雅, 河原 英紀, 西浦 敬信: 基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法, 電子情報通信学会 論文誌 D, vol. J93-D, no. 2, pp. 109–117, 2010.