J-018

発話音声から受ける要素感覚を決める音響特徴量の評価指標

A Proposal of Evaluation Index of Acoustic Feature for Deciding Element Sense Received from Speech

西田 悠 † Haruka Nishida　浅田 龍星 † Ryusei Asada　中平 勝子 † Katsuko T. Nakahira　北島 宗雄 † Muneo Kitajima

1　はじめに

本稿ではコミュニケーション場の定量評価を可能とするため，コミュニケーション場における印象評価に繋がる要素感覚を規定する音響特徴量を算出し，指標化された音響特徴量の決定過程を，印象を特定できる場面に適用することで，各要素感覚に対する音響特徴量の範囲を特定し，指標化する．

コミュニケーション場を，ある基準に従って表現することにより，各種研修で取り入れられているディベートやアクティブラーニングにおけるコミュニケーション場の客観的記述が可能となることが見込まれる．すなわち，記述されたコミュニケーション場の状態に対する分析手法を与えることで，発展的には，コミュニケーション場で語られる内容のみならず，それがどの様な状態で発せられたかを含めた分析が可能となる．また，ディベートやアクティブラーニングの場の状態分析・評価形成に大きく寄与できると考えられる．

そうしたプロセスを経て，最終的には職場における社員教育をはじめ，各種教育場面でのディベートにおける成果計測への適用も期待される．本稿は，その第一歩としての指標化に対する基礎研究と位置づけ，システムに必要な機能の一部の実装と特徴量の指標化を行った．

各特徴量は，発話音声をケプストラム法や音素セグメンテーションアルゴリズム [1] を用いて抽出する．

2　音響特徴量抽出

浅田ら [2] によると，発話印象のモデルは，6 種の音響特徴量，5 種の要素感覚，5 種の発話印象の 3 階層で表現可能である．

本稿で開発される要素感覚に対する音響特徴量の指標は図 1 の様な流れの中で用いることを想定している．コミュニケーション場で収集される音声情報から音響特徴量を算出し，これらを本稿で規定する手順でレベル化する．それを元に要素感覚と結びつけるための指標化を行い，定量表現された要素感覚を用いて発話印象を表現する．各音響特徴量と要素感覚の対応付けはある程度決まっているが，各要素感覚と発話印象の対応付けははっきりとは分かっていない．抽出する音響特徴量は，振幅スペクトル，ピッチ，モーラ数，ポーズ比で， これらに対応する要素感覚は，声量・抑揚・明瞭性・ポーズ長・話速である．

特徴量の抽出は次のように行った．処理は 1 発話を単位とした．なお，1 発話は，一人の話者が話し始めてから話し終え，次の話者が話し始めるまでを 1 発話とする．また発話の仕方によって受ける感覚が調査の主眼であるため，相槌は発話とせず解析から除外した．

まず，1 発話音声のノイズを除去した後に音声区間を検出し，1 発話音声の長さと無音区間の長さの比をポーズ比とする．そして母音のあるフレーム (VF) を特定し，1 発話内の文字数を得る．これを 1 発話の時間で割った 1 秒間の文字数をモーラ数とする．次に VF 毎にケプストラム分析を行い，振幅スペクトルとピッチを求める．

音声区間検出，モーラ数抽出には音素セグメンテーションのアルゴリズムを用いた．発話音声をフレーム分割し，各フレームにおいて有声音検出パラメータ v を求める．フレーム長は 10 [msec] で，オーバーラップはしない．各フレームのメル対数スペクトルを求める．窓関数はフレーム長のブラックマン窓で，メルフィルタバンクのチャンネル数は 20 である．メル対数スペクトルの 4 から 17 次元目の数値の平均値を v とし，v の値が閾値を下回っているフレームを無音とする．閾値は v を昇順ソートした際の第一四分位数とした．この範囲は基本周波数帯域部分で，おおよそ 75.3 [Hz] から 323.3 [Hz] である．

モーラ数の抽出には，メル周波数ケプストラム係数のデルタパラメータを用いる．各フレームの零次 MFCC を抽出し，時系列データを作成する．デルタパラメータを求めたいフレームと前後の 2 フレームを含めた 5 フレームから，デルタパラメータを計算する．得られたデータ列における極大値の中で，閾値を越えているデータの個数を音素数とし，これを発話時間で割ったものをモーラ数とする．閾値は，データ列を昇順ソートした際の第一四分位数とした．

ピッチの抽出には，ケプストラム法を用いる．各 VF の振幅スペクトル，ピッチを抽出するために，各 VF の 25 次以上の高次ケフレンシ領域を解析対象とした．ピッチを求める際のフレーム数は，情報量を増やすために VF の前後 2 フレームを含めた 5 フレームとし，FFT 点数は 2048 点とした．一つの会話音声において 3 人が何回かずつ発話しており，各発話においてモーラ数分の特徴量が求められ，これらの最大値・第三四分位数・中央値・第一四分位数・最小値を抽出する．振幅スペクトルは，各音素のスペクトルの中の基本周波数帯域部分を見て，これの中央値をその音素における代表の振幅値とした．ここから，他の特徴量と同様に最大値などを抽出する．

西崎ら [3] が示している特徴量の範囲などを参考に，各要素感覚を決める特徴量を以下に示すように指標化する．

- 振幅平均 [dB]：[60, 70]
- ピッチ平均 [Hz]：[135, 140]
- ピッチ標準偏差 [Hz]：[28, 30]
- ポーズ比 [%]：[10, 15]
- モーラ数 [モーラ/秒]：[7.6, 8.1]

† 長岡技術科学大学

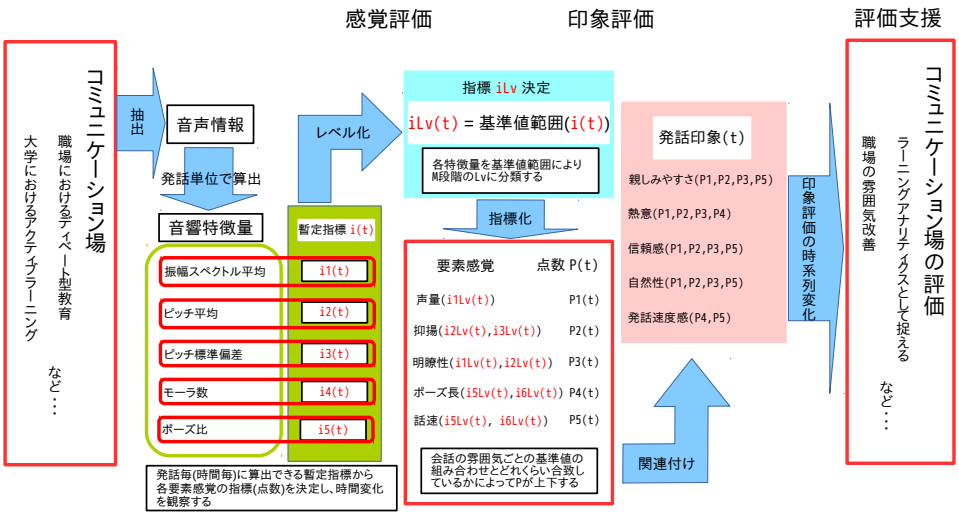

図 1　システムの構想図

各特徴量の範囲はそれぞれ，一般的な会話における振幅値の範囲，声が明瞭と感じた場合のピッチ平均，抑揚がついていると感じた場合のピッチ標準偏差，適切にポーズを入れていると感じた場合のポーズ比，話速が丁度良いと感じた場合のモーラ数である．抽出された特徴量の中央値が下限未満の場合，これを $Lv1$ とする．範囲内に収まっている場合，これを $Lv2$ とする．また上限より大きい場合，これを $Lv3$ とする．これらのレベル値の組合せにより，特定の雰囲気における会話を示す特徴量の指標を決定する．特徴量の抽出は 1 発話単位で行うことで，指標の時間変化を観察し，その場の評価を行う．

3　音響特徴量-要素感覚の関連付けの例

解析に用いた音声は，千葉大学 3 人会話コーパスの音声である．サンプリング周波数は 16000 [Hz]，量子化 bit 数は 16 [bit] の WAVE 形式である．ここで一般的なボイスレコーダーなどによる録音を想定し，サンプリング周波数を 44100 [Hz] に変更した．対象として，コーパス音声の中から，雰囲気が明るく，円滑に会話が進んでいるという感覚を受ける会話を選択した．音声の収録は 3 人の女性の友人同士が正三角形の形で向かい合い，互いの距離は 1.5 [m] ほど離れていて，リラックスした雰囲気の中で行われている．

表 1 に抽出された特徴量および判定されたレベルを示す．結果として，振幅平均は $Lv2$，ピッチ平均は $Lv3$，ピッチ標準偏差は $Lv1$，ポーズ比は $Lv3$，モーラ数は $Lv2$ となった．

会話内では笑いながら話している場面が多く，講義などと異なりほとんど一定のトーンで話しているわけではない．よって，声量が大きく，周波数は高くなると言えるので，ピッチ平均が大きくなっていると考えられる．また，全て女性が話していることも原因と考えられる．コーパスの録音環境から，互いの声はしっかり聞き取れる位置で明瞭に聞こえ，丁度良い声量であると考えられる．振幅平均は範囲内に収まっていて，妥当であると言える．ピッチ標準偏差は小さくなっているが，抑揚があまりついていなくても円滑なコミュニケーションを行うには問題ないと考えられる．ポーズ比は大きくなっているが，講義などと異なり，ポーズが多少長くとも不満を覚える話し方にはならないと考える．

表 1　抽出された特徴量の例

特徴量	中央値	判定されたレベル
振幅平均 [dB]	64.8	$Lv2$
ピッチ平均 [Hz]	158.3	$Lv3$
ピッチ標準偏差 [Hz]	15.5	$Lv1$
ポーズ比 [%]	26.0	$Lv3$
モーラ数 [モーラ/秒]	7.6	$Lv2$

各特徴量と各要素感覚の対応付けから，以下の結果を得る．

- 声量：振幅平均 $Lv2$
- 抑揚：ピッチ平均 $Lv3$ ピッチ標準偏差 $Lv1$
- 明瞭性：振幅平均 $Lv2$，ピッチ平均 $Lv3$
- ポーズ長：モーラ数 $Lv2$，ポーズ比 $Lv3$
- 話速：モーラ数 $Lv2$，ポーズ比 $Lv3$

4　まとめと今後の課題

本研究では円滑なコミュニケーション支援システムの開発を目指し，音声から受ける要素感覚を音響特徴量から特定するための指標を決定した．

今後の課題として，他の印象を受けるコミュニケーション場における指標を調査することで，実際のディベート教育やアクティブラーニングに役立てていくことができると考える．

また，話者検出や語尾音素の特定，相槌検出を加えるなどして，雰囲気改善を支援してコミュニケーションを円滑に進めるための支援システムの構築を目指していく．

参考文献

[1] 今井聖．音声認識．共立出版株式会社, 1995.
[2] 浅田龍星, 西田悠, 中平勝子, 北島宗雄．合成音声を利用した好印象発話モデルの構築．FIT2017(印刷中), 9 2017.
[3] 西崎博光, 関口芳廣．教員の話し方改善支援システムの開発に向けた講義音声の特徴分析．日本教育工学会論文誌, Vol. 34, No. 3, pp. 171–179, 2010.