ニューラルネットワークを用いた感情音声の認識

山崎開人† 金子正人†

本研究では，あたかも人間を相手にしているような HMI(Human Machine Interface)を目指している．実験を行う中で，画像認識などでよく使用される特徴量の変化量を感情音声に適用することで，認識率を上昇させることができるのではないかと考えた．本研究では従来モデル，変化量を用いたモデル，両者を用いた混合モデルの３種を用いて，学習済み・未学習データにおける評価を行った．その実験の結果，混合モデルが最も良い結果を出していることを確認した．

Emotional speech recognition using neural networks

Kaito Yamasaki† and Masato Kaneko†

This research focuses on an HMI (Human Machine Interface) that feels as if it is Dialoguing with a human. During the experiments, we thought that it might be possible to improve the recognition rate by applying the difference of amount in feature often used in image recognition to emotional speech. In this research, we evaluated three types of models using trained and untrained data, which is a conventional model, a difference model, and a mixed model using both. As a result of the experiment, the mixed model recorded the best result was confirmed.

1. はじめに

我々はあたかも人間を相手にしているような HMI(Human Machine Interface)を目指して，機械学習を用いた感情音声の認識実験を行ってきた．時間やピッチ，振幅長を用いた特徴量を基本 8 パラメータとし，精度向上に取り組んだ．その中で，脳波などを用いて特徴量を増加させる手法の研究では，認識率を大きく上昇させるような結果を出せなかった．そこで，新たな特徴量として，画像認識などでよく使用される特徴量の変化量を感情音声認識に適用することで，認識率を上昇させることができるのではないかと考えた．

本研究は，感情音声の認識で使用できる新たな特徴量を提案し，その有効性を検証するものである．

2. 提案手法

前述したように，新たな特徴量として変化量を使用して実験を行う．先行研究では，感情音声の時間，ピッチ，振幅長を用いた８パラメータを分析している．このパラメータを使用し，平静時との変化量を各感情で求める．

3. 感情の抽出

発話実験と聴取実験の 2 種の実験を行い，音声から感情を抽出する．発話実験で感情を付加した音声サンプルを録音し，聴取実験で感情の他者評価を行う．本研究では，発話者が付加した感情ではなく，聴取実験での評価を正解として教師データを作成する．

4. 発話実験

認識実験のデータセットに必要な，感情を付加した音声の発話実験を行う．本研究では，短文「あぁ，雪だ」に「平静」,「喜び」,「悲しみ」,「怒り」,「嫌々」,「驚き」の 6 感情を付加した音声を発話する．発話実験では，録音機器が発話者の細かな動作を記録するため，録音を開始して 2 秒後に発話させ，発話後も 2 秒経過してから録音を停止する．これを 6 感情で行う． 本研究では，15 名の協力を得て，発声前後をトリミングした 90 個の音声サンプルを用意した．15 名の被験者の内訳は，男性が 10 名，女性が 5 名である．例として，トリミングを行った男性 1 名の音声サンプルを図 1 に示す．

† 日本大学工学部
College of Engineering, Nihon University

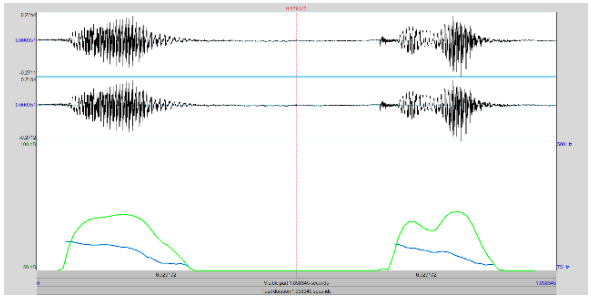

図１.「あぁ，雪だ」の音声サンプル（平静）

5. データ拡張

音声編集ソフトを使用し，発話実験で取得した音声サンプルのデータ拡張を行う．聞々ハヤえもん[1]と WavePad[2]の２種の音声編集ソフトを使い，時間，音程，最大振幅長の３項目をそれぞれ変化させ，１つの音声サンプルから７つの拡張データを作成した．すべての項目においてデータ拡張を行うと，元の音声サンプルに比べ，後述するパラメータが該当項目以外も少し変化する．これにより，単一の項目の変化でも，異なるデータの拡張音声を作成することができる．

5.1. 時間

発話実験で取得した音声サンプルを元に，再生時間を 1.1 倍にした音声と 0.9 にした音声を作成した．再生時間を 1.1 倍にした音声は，通常の音声に比べ，少しゆっくりと発声しているように聞こえる．逆に，0.9 倍にした音声は，少し早口に聞こえる．

5.2. 音程

発話実験で取得した音声サンプルを元に音程を，半音下げ，半音上げ，全音上げの音声を作成した．音程を下げた音声は，通常に比べて声が太く聞こえる．逆に，音程を上げた音声は，声が細く聞こえる．

5.3. 最大振幅数

発話実験で取得した音声サンプルを元に，最大振幅長を 5db 上げた音声と，5db 下げた音声を作成した．例として，図１の音声を 5db 上げた音声を図２に示す．図の上部が音声波形，下部の青い線がピッチ，緑の線が最大振幅長を表す．図１と比較すると緑の線が大きく盛り上がっていることがわかる．他の項目においても，図２のような変化が該当項目で起きている．

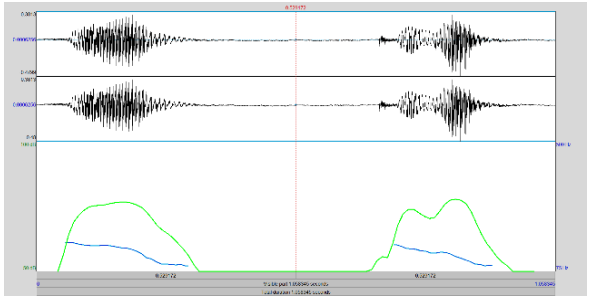

図２． ５db 上げた音声サンプル（平静）

6. 聴取実験

発話実験で録音した音声サンプルに，正しく感情が付加されているか検証するため，聴取実験を行う．聴取実験では，被験者に各音声サンプルの「平静」を聞かせた後に，他の感情音声をランダムに聞かせ，印象値を記録させる．形式は，感情の強さの印象を 0（全く感じない）から 10（最も感じられる）の 11 段階とした．複数回答可の条件で，録音したすべての音声サンプルの評価を行った．例として，男性１名の音声サンプルに対する各感情の評価値の合計を表１に示す．

表１．音声サンプルの評価

		被験者の評価					
		平静	喜び	悲しみ	怒り	嫌々	驚き
音声サンプル	平静	20	0	2	0	0	0
	喜び	2	15	0	0	0	3
	悲しみ	7	0	9	0	3	6
	怒り	0	0	1	20	2	0
	嫌々	0	0	7	2	17	0
	驚き	0	2	0	0	0	20

7. 音声解析

音声解析プログラムを使用し，すべての音声サンプルのパラメータ解析を行う．本

研究では，従来使用していた音声解析ソフトではなく，Python ライブラリである Librosa や PyWorld を使用し，パラメータ解析を行った．パラメータ解析では，8 項目（「時間長」,「平均ピッチ」,「最大ピッチ」,「最小ピッチ」,「始まりのピッチ」,「終わりのピッチ」,「最大振幅」,「発声時間」）の解析を行う．

前述のように男性 10 名，女性 5 名の計 15 名に協力を得て発話実験を行った． 男性と女性の全感情のピッチパラメータの平均を比較するグラフを図 3 に示す．グラフの青は男性，赤は女性を表す．時間長，発声時間，最大振幅長は男女ともに大きな差はなかったが, ピッチに関してはすべての項目において女性の方が高い値を記録した．最大ピッチは 1.5 倍，平均ピッチと最小ピッチは 1.7 倍，始まりのピッチと終わりのピッチは 1.8 倍高い結果となった．そのため，本研究は男女それぞれで実験を行う．

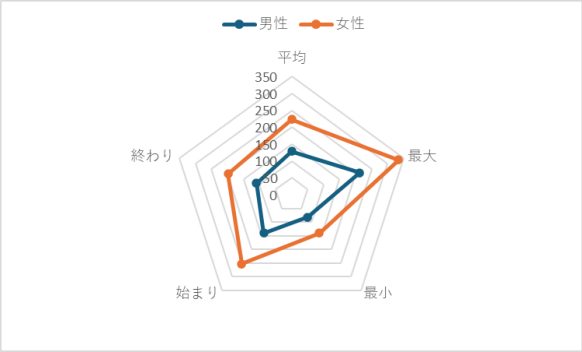

図 3 . ピッチパラメータ（6 感情の平均）

8. 学習・教師データの作成

データの正規化または標準化を行い，音声解析により得たパラメータを学習データ，聴取実験で得た評価を教師データとして使用する．式（1）は特徴量の正規化の計算式を表す．人の声の周波数の範囲は概ね 100～1000Hz である．x の最小値を 100Hz，最大値を 1000Hz と固定して正規化することで，学習データを作成することは可能だが，本研究では最小ピッチが 100Hz 未満のデータが多数あったため，パラメータ解析で得たデータの最大値と最小値を使用して正規化を行う．教師データについては，前述した音声サンプルの評価を正規化することで作成する．音声サンプルの評価では，ほぼすべてのデータが 0 を 1 つ以上含んでいたため，すべてのデータにおいて，最小値を 0 として正規化を行った．

$$g(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \cdots \quad (1)$$

式（2）はパラメータの標準化の計算式を表す. 本研究の提案手法である変化量は，正規化ではなく標準化を用いる．

$$f(x) = \frac{x - \mu}{\sigma} \quad \cdots \quad (2)$$

拡張した音声サンプルの内，時間と最大振幅長の 2 項目は，元の音声サンプルと聴き比べると大きな違いを感じることはなかった．そのため，元の音声サンプルの教師データをそのまま使用する．

前述したように，本研究では男性と女性のそれぞれで実験を行い，結果を出力させる．そのため，学習・教師データは男女の 2 種類作成する．

9. 認識実験

9.1 実験方法

本研究では，変化量の有効性を検証するため，従来の 8 パラメータを用いたモデル（以下 base とする），変化量のみを用いたモデル（以下 diff only とする），両者を用いたモデル（以下 base+diff とする）で実験を行い，比較検討する．

9.2 認識距離

認識距離とは，聴取結果と認識結果の差をそれぞれ求め，それらをすべて合計したものである．認識距離を D，聴取結果を Tn，認識結果を Rn，感情数を N とすると，式（3）のように表せる．認識距離の値が小さいほど人間の感情認識に近い結果と判断される．

$$D = \sum_{n=1}^{N} |T_n - R_n| \quad \cdots \quad (3)$$

10. 結果と考察

学習済みデータにおける男性の認識結果を図 4，女性の認識結果を図 5 に示す．グ

ラフの縦は認識距離を表し，横は各感情を表す．図４に着目すると，すべての感情において，base より diff only，base+diff の方が良い結果となった．また，平静以外の５感情では base+diff が最小の値を記録した．特に，「平静」，「喜び」，「悲しみ」，「驚き」の４感情では認識距離が大きく改善されている．しかし，「平静」に関してはひとつ問題点がある．発話者それぞれで平静を基準に変化量を求めたが，平静と平静の変化量は０である．つまり，異なる発話者だが同じ値を記録しているような状態になっている．そのため，このような結果になったと考えている．この同一の値を全発話者の平静の基準値として考えることはできるが，標準化する際に平均値や標準偏差を求める必要があるため，汎用的な数値ではない．この問題は男女，学習済み，未学習のすべての実験結果おいて共通する課題である．しかし，これは「平静」だけの問題であり，他の感情は純粋な変化量を表している．そのため，すべての感情において最小値を記録した変化量を含むデータセットは，新たな特徴量として有効である．

図５に着目すると，男性と同様にすべての感情において，base より diff only，base+diff の方が良い結果となった．「平静」，「喜び」，「悲しみ」の３感情では，base+diff が最小値を記録した．一方で，「怒り」，「嫌々」，「驚き」の３感情では，diff only が最小値を記録したため，base+diff と diff only の優劣を着けることはできない結果となった．すべての感情において変化量を用いたデータセットが最小値を記録したため，変化量を用いることは有効である．

未学習データにおける男性の認識結果を図６，女性の認識結果を図７に示す．図６に着目すると，「平静」，「怒り」，「嫌々」，「驚き」の４感情において base+diff が最小値を記録した．「喜び」，「悲しみ」では，diff only が最小値を記録した．学習済みデータのときとは違い，「嫌々」では diff only が一番悪い結果を記録した．また，すべての感情，すべてのデータセットにおいても，学習済みのときと比べ，値が倍近く上昇している結果が散見される．しかし，すべての感情において変化量を用いたデータセットが最小値を記録したことを踏まえると，変化量を用いることは有効である．

図７に着目すると，「平静」，「悲しみ」では diff only が最小値を記録した．「喜び」では base，「怒り」，「嫌々」，「驚き」では base+diff が最小値を記録した．学習済みデータのときとは違い，「喜び」では base+diff が一番悪い結果を記録した．男性の未学習と同様で，すべての感情，すべてのデータセットにおいて，学習済みのときと比べ，値が倍近く上昇している結果が散見される．しかし，半数以上の感情において変化量を用いたデータセットが最小値を記録したことを踏まえると，変化量を用いることは有効である．

以上が，本研究におけるすべての実験結果である．全体を通して最小値を最も多く記録したのは base+diff だった．これは入力するパラメータの数による差であると考えられる．従来の８パラメータに加え，その変化量の８パラメータの計 16 あるため，他のデータセットに比べると倍である．しかし，base+diff を除いても，ほぼすべての実験において base より diff only の方がよかったため，変化量は有効である．学習済みと未学習を比較した際，ほとんどの実験結果において値が倍近く上昇していることに関しては，被験者の少なさと感情の判別が困難な音声サンプルが影響していると考えられる．

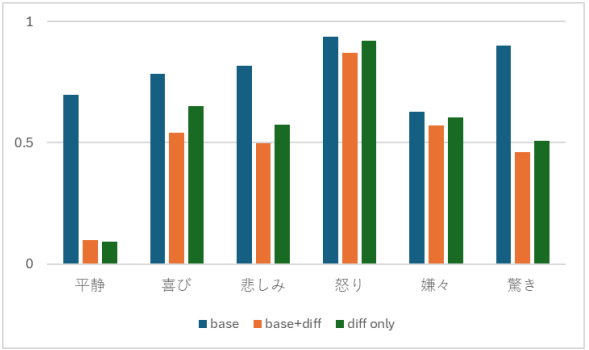

図４．男性の認識結果（学習済み）

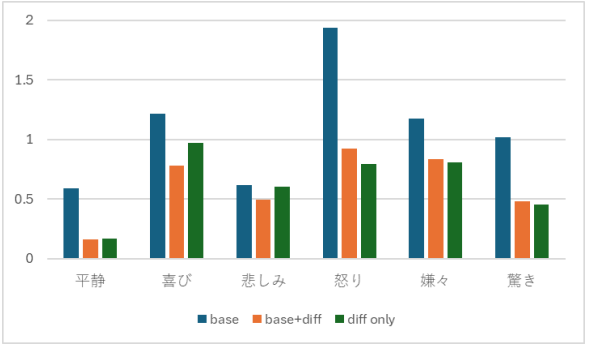

図５．女性の認識結果（学習済み）

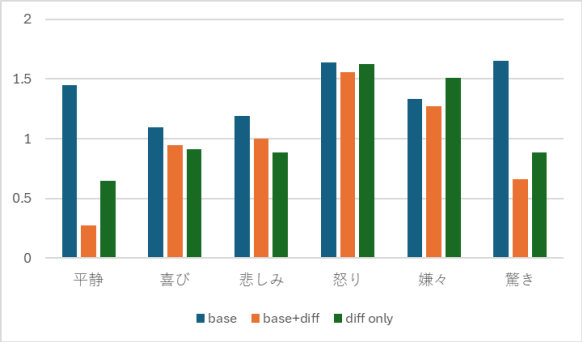

図６．男性の認識結果（未学習）

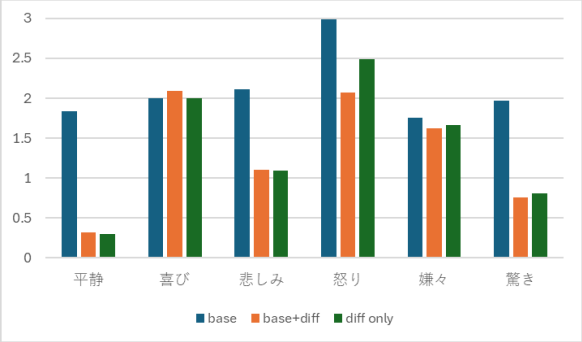

図７．女性の認識結果（未学習）

11. おわりに

本研究では，感情音声の認識における変化量の有効性を検証した．結果として，学習済み･未学習データにおいて，従来の 8 パラメータよりも優れた結果を確認できた．今回は平静との純粋な変化量を使用したが，変化の大きさに着目する場合は絶対値や変化量の２乗を使用することができる．このような変化量を応用したパラメータを使用することで，同様に精度が向上する可能性があると考える．

今後の課題としては，前述した「平静」の問題に取り組む必要がある．また，聴取実験の最適化にも取り組む必要がある．データ拡張の項目を増やせば増やすほど，聴取実験にかかる時間は大きく増加するため，より低コストな方法を検討しなければならない．数千の音声サンプルをデータ拡張せずに用意した場合であっても，同様の問題に直面するため，聴取実験の最適化は重要な課題であると考える．

参考文献

1) 聞々ハヤえもん，https://hayaemon.jp/
2) WavePad，https://www.nchsoftware.com/
3) 横井，金子，武内，藤本：“音声に含まれる感情情報の認識に関する研究”,FIT2005