E-050

乳児音声区間の検出と感情認識への応用

Detection of Baby Voice and its Application to Emotion Recognition

山本 翔太†＊　吉冨 康成††　田伏 正佳††　櫛田 康ǂ
Shota Yamamoto　Yasunari Yoshitomi　Masayoshi Tabuse　Kou Kushida

1. まえがき

乳児の育児には手がかかり，また，ひとときも目を離せないことも多く，その負担は大きい．他方，日本では近年少子化が進み，その対策が行われつつある[1,2].

自宅内の乳児から離れた場所で，家事をこなしたり，趣味に興じたりするため，無線マイクを乳児の近くに置くとする．このような対策をとっても，乳児から離れてはいるものの，マイクから送られてくる音声に耳を傾け続ける必要があるので，育児ストレスの軽減は困難と考えられる．

そこで，母親などの育児ストレスの軽減などを目指して，周波数解析をもとにした乳児音声の解析・理解の研究[3 - 8]が行われてきたが，録音された音声波形を見て対象とする乳児音声をマニュアル操作で採取する必要があるため，これらの研究を育児支援システムの開発に生かすのは容易でない.

著者らは，育児負担の低減が少子化対策の一助になると考え，また，母親らが育児ストレスを軽減した状況で，自宅内の乳児から離れた場所で，家事をこなしたり，趣味に興じたりすることを支援することを目指して，育児支援システムの研究開発に着手した．本研究では，Julius[9]による音声認識と基本周波数の短時間での変化の特徴を用いて，連続音声から乳児音声区間を検出する方法を開発した.

そして，本法の応用例として，検出した区間の音声に対し，著者らが既報の研究[8]で用いた，FFT による周波数解析により見出した特徴量をもとに，主成分分析を用いて感情特徴のパターン認識を行い，本法の有効性を検証した.

2. 幼児の言語獲得について

乳幼児の言語獲得は，学習を中心とする知覚と生成の二つの側面があり[10]，母語の影響を強く受け生得的なものである．聴覚は胎生 7 ヶ月前後で完成し，胎児段階から外界の音声を聴き学習している．生後 200 日ごろから母語への選択的注意が起こることが指摘されている[10]．乳児は生後 6 ヶ月を過ぎたころに喃語発声期を迎える．喃語発声初期は，発声器官が未熟で子音の音質はあまり良くないが，成長に伴って明瞭な子音が現れ，音節の組み合わせも，月齢が上がるにつれて複雑になる．喃語発声後期には，初語が出現し，10 ヶ月児で発声頻度の分布が母語と一致することが報告されている[11]．乳児では，平均的な発声時間は二語文が現れるようになった後には，直線的に増加していく[12]．やがて 1 歳半を過ぎ二語文の時期になると，文の長さの伸びに伴って，再び 1 秒以上の長い発声が増えていく．実際は，発達過程と呼応した回帰的な性質をもつ発声時間の伸びが存在する[13].

†京都府立大学，Kyoto Prefectural University
††京都府立大学大学院，Graduate School of Kyoto Prefectural University
ǂ京都府立田辺高等学校, Tanabe Senior High School
＊株式会社　日本公文教育研究会, KUMON EDUCATIONAL JAPAN CO., LTD.（現在）

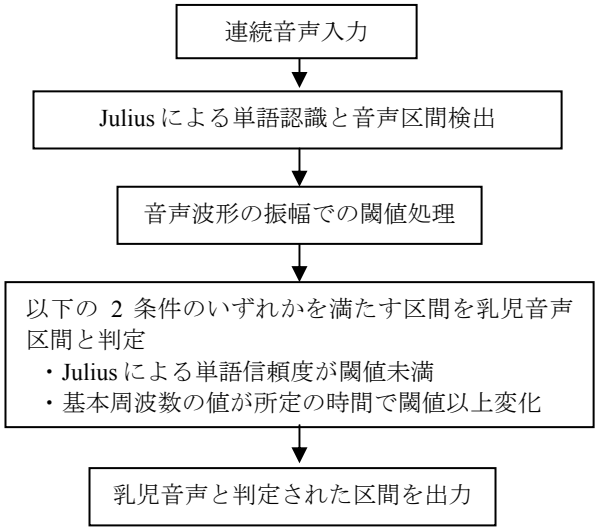

図 1　乳児音声区間の検出処理のフロー

本研究では，喃語発声期以前(生後 6 ヶ月未満)の乳児を対象とした．従って，本研究では，「対象となる乳児が単語を発声することはない」と想定して，乳児音声区間を検出する．実際に本研究で対象とした乳児の音声は，喃語に近いものではなく，「泣き声」である.

3. 方法

本法の処理概要を図 1 に示す．以下で，各処理について説明する.

3.1 Julius による単語認識と音声区間の検出

音声区間検出では，入力ストリームに対して短時間ごとの特徴から音声区間の開始・終了を検出し，それをもとに認識単位の切り出しおよび発話単位の区切りを行う．本法では，Julius を用いて，振幅と零交差に基づく入力検知の方法を用いた．そして，Julius が，対象区間について何らかの単語を認識した場合(Julius における表記で silB, sp, silE, 以外を出力した場合)，その区間の音声波形を以降の処理に供する.

3.2 音声波形の振幅での閾値処理

3.1 節記載の方法で音声区間と判定された区間の中に，雑音のみの区間も含まれる．これは，音声収録時の生活音が主な原因と考えられる．本法では，振幅に閾値をもうけ，その値を超える振幅をもつか否かで雑音のみの区間か否かを判定する．また，その際，閾値を実験的に決定し，無音区間における音声波形の振幅の平均 m，標準偏差 σ を用いて，$m+2\sigma$ を閾値とした.

3.1 節記載の方法で音声区間と判定され，かつ上記の振幅による閾値処理により，雑音と判定されなかった区間を，乳児音声区間候補とする.

3.3 乳児音声区間の検出

乳児音声区間候補のうち，以下の 2 条件のいずれかを満たすものを乳児音声区間と判定する．

条件 1．Julius による単語信頼度の値が閾値未満

条件 2．基本周波数の値が所定の時間で閾値以上変化

3.3.1 単語信頼度による判定

Julius では単語認識結果に対して，信頼度を出力できる．単語信頼度は事後確率に基づいて算出され，0 から 1 の値を取る．1 に近いほど，競合候補に比べて尤度の差が大きかったことを表し，認識結果として「信頼度が高い」と解釈できる．また，0 に近い場合，似た確率を持つ多くの競合候補が存在したことを表し，認識結果として「信頼度が低い」と解釈できる．

本研究で，成人音声に比べ乳児音声の方が，単語信頼度の値が低い傾向があった．そこで，閾値を設定し，この現象を乳児音声区間の検出に用いた．また，この単語信頼度の閾値として，予め採取した音声において，乳児音声の頻度の累積から成人音声の頻度の累積を引いた値が最も高くなった単語信頼度の値を閾値として用い，乳児音声区間候補のうち，単語信頼度が閾値未満のものを乳児音声区間と判定した．

3.3.2 基本周波数の短時間での変化による判定

本法では，自己相関を用いて基本周波数を測定する方法[14]を用いた．そして，本研究で，乳児音声の基本周波数の値が，短時間で大きく変化することが観察された(図2；1.5～1.8s, 6.6～6.8s の各区間)．そこで，予め採取した音声をもとに，測定時間間隔と基本周波数変化の閾値を定め，乳児音声区間候補のうち，基本周波数が閾値以上変化する区間を乳児音声区間と判定した．

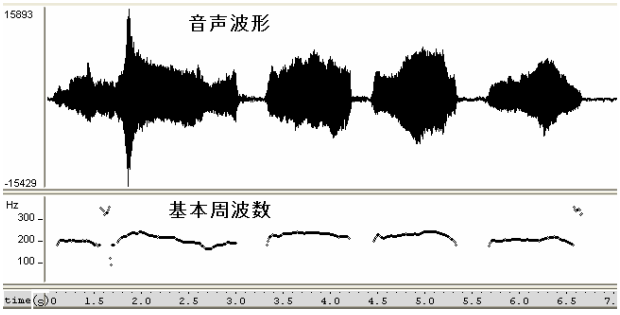

図 2　乳児音声の基本周波数の急激な変化

4. 評価実験

4.1 方法

4.1.1 音声の収録

乳児音声の収録は，有線または無線マイクとノート型 PC により行った．入力された音声は，wave 形式のファイルとして保存される．このときの音声データは，PCM フォーマット，16 kHz，16 ビットのモノラル音源とした．

自宅で収録した，月齢 1.5 ヶ月の乳児の，57 の連続音声(1 連続音声につき 10～12 発声を含む)と，保育園で収録し，乳児音声と成人音声が混在する 3 連続音声(1 連続音声につき 8～21 発声を含む)(「混合①」と表記)と 5 連続音声(1 発声につき 17～21 発声を含む)(「混合②」と表記)と，大学の研究室で収録した，成人 7 人の 7 連続音声(日本名「taro」,1 連続音声につき 14～22 発声を含む)(文献[15]記載の音声のうち「無表情」のもの)，を用いて，本法の評価実験を行った．なお，開発環境としては，PC: DELL OPTIPLEX GX260(CPU:INTEL Pentium 4　2.4 GHz, メモリ:512 MB)，OS: Windows XP, プログラム言語:Microsoft Visual C++ 6.0, を用いた．単語認識には Julius4.0 を用いた．

4.1.2 性能評価方法

まず，月齢 1.5 ヶ月の乳児と成人 7 人の音声データに対し，前半部を学習データとし，3 章記載の方法で，振幅による足切の閾値，および Julius による単語信頼度の閾値

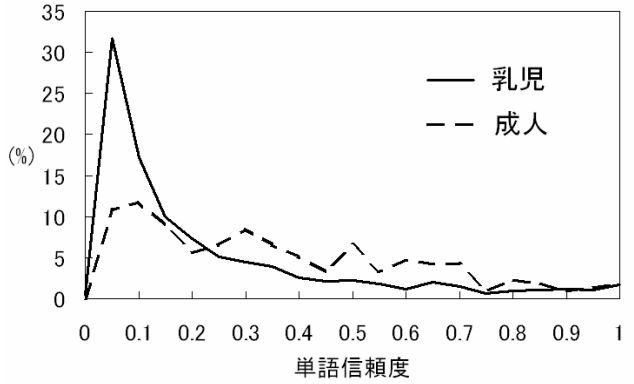

図 3　乳児と成人の単語信頼度の分布

0.15 を求めた．単語信頼度の乳児と成人 7 人の分布を図 3 に示す．また，基本周波数の急激な変化の有無を判断する条件を実験的決定し，時間間隔を 0.1s，変化量の閾値を 150Hz とした．

また，対象とした音声データは，Julius により検出された全区間に対し，波形の目視と聴音確認により乳児音声か否かを判断した．そして，これにより得られた乳児音声区間と本法により検出された区間とを比較した．その際，評価を行う尺度として検出率と誤検出率を以下のように定義した．

まず， 図 4 において，A を「乳児音声区間」の集合とし，B を本法による「検出区間」の集合とする．さらに，A∩B の集合を C とし，全区間の集合を D とする．その時，N(X)を集合 X に含まれる要素の数として，N(C) / N(A) つまり『乳児音声区間のうち，検出されたものの割合』を検出率とし，(N(B) - N(C)) / (N(D) - N(A))　つまり，『乳児音声区間以外のうち，本法により検出されたものの割合』を誤検出率とした．

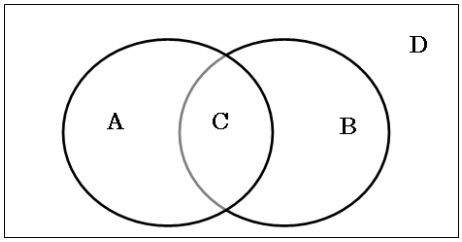

図 4　検出率と誤検出率の関係

本法の応用として行った感情認識実験では，月齢 1.5 ヶ月の乳児の音声に対し，乳児音声区間を本法で検出したものと手動で検出したものをそれぞれ感情認識し，結果を比較した．感情パターンは，「不快」，「空腹」，「眠気」の 3 つとし，乳児の保護者が 3 感情のいずれであるかを音声採取時に判定しておいた．また，音声データは，前半部を学習データ，後半部を認識データとした．

感情認識の処理フローを図 5 に示す．まず，学習用の音声に 32 次元 FFT を施し，各周波数成分(パワー)の値から無音区間での値を引き，得られた特徴量を元に学習データに対し主成分分析を行い，累積寄与率 80%を超えるまでの成分を用いて音声モデルを構築する．そして，認識用の音声に対しても同様に，32 次元 FFT を施し，各周波数成分(パワー)の値から無音区間での値を引き，得られた

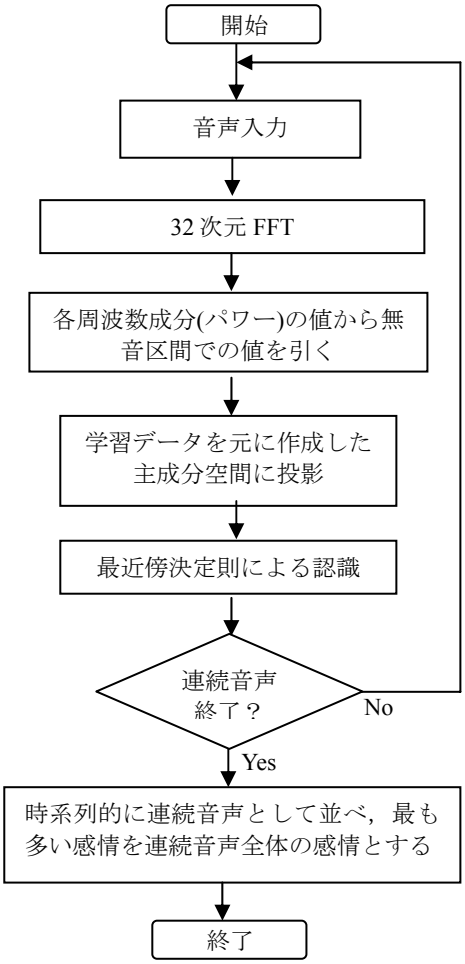

図 5　乳児音声からの感情認識のフロー

特徴量を，学習データを元に作成した主成分空間に投影後，ユークリッド距離による最近傍決定則を用いて認識を行った．各区間に対する認識結果を時系列として並べなおし，最も多い感情のクラスをその連続音声全体の認識結果とする．また，その際，複数の種類の感情クラスにおいて各区間の認識結果の個数が等しい場合，認識データに 2 番目に近い学習データのクラスを加えて，最も多い感情のクラスをその連続音声全体の認識結果とする(以降，同数が続いた場合，3 番目，4 番目も考慮する)．

表 1 に説明のための例を示す．この例では，連続音声として並べなおした結果，最も多いクラスである「空腹」を連続音声全体の認識結果とする．

4.2　結果と考察

4.2.1　乳児音声検出

表 2 に，乳児音声検出結果を示す．検出率は，65.0～69.4%と比較的近い値となっているのに対し，誤検出率は 2.0～34.1%と幅がある．「混合①」の場合，騒音(室外の車の音や乳児をあやすために用いたカタカタの音など)の非常に多いものであったため誤検出が多かったと考えられる．騒音区間は，Julius による単語信頼度の値が低くなり誤検出を起こしやすい．また，「混合①」および「混合②」において，成人が乳児をあやす際の発声も非言語となっているため，単語信頼度の値が低くなり誤検出を

表 1　連続音声の感情認識

データ番号	不快	空腹	眠気	判定
1			○	空腹
2	○			
3		○		
4			○	
5		○		
6	○			
7		○		
8		○		

表 2　乳児音声検出結果

	乳児	成人	混合①	混合②
Julius による検出区間数	1172	333	186	239
乳児音声区間数	569	0	60	126
本法による検出区間数 (乳児音声区間数)	407 (395)	66 (0)	82 (39)	97 (94)
検出率	69.4%	——	65%	66.2%
誤検出率	2.0%	19.8%	34.1%	2.7%

起こしやすかった．「成人」の場合(「taro」と発声)の誤検出の原因も単語信頼度を 1 つ判定基準にしていることによる．図 3 に示すように，成人の単語発声の場合でも，単語信頼度の値が低くなる場合がある．

育児負担を減らす必要性が高いのは，自宅で育児を行う場合と考えられる．表 2 では，「乳児」の場合が該当する．この場合の検出率 69.4%，誤検出率 2.0%，という結果は，本法が有効であることを示唆していると考えられる．

自宅内の乳児から離れた場所で，家事をこなしたり，趣味に興じたりする場合，無線マイクを乳児の近くにおいて，マイクから送られてくる音に耳を傾け続けるのと，本法を利用して乳児が泣いた時だけ通知するシステムを利用するのを比較すると，後者の方が育児ストレスは少ないと考えられる．また，後者の場合，乳児が泣いた時に乳児の映像を送るシステムを構築することで，乳児からある程度離れた場所に居ることも可能になると考えられる．

4.2.2　感情認識への応用

次に，本法または手動で検出した乳児音声に対し，感情認識を行った結果を表 3 に示す．3 感情の平均認識率は本法では 62.1%，手動では 66.6%となった．このため，感情

認識に供するに際し，本法による乳児音声区間の検出は，手動での検出とほぼ同等の性能があったと考えられる．

「不快」感情は他の感情に比べ，本法，手動とも認識率が低い．これは，「不快」感情の要因が多岐に渡っているため，乳児の泣き方に多くのパターンがあることによると考えられる．また，誤認識を起こした認識データの特徴ベクトルと 2 番目にベクトル距離の近い学習データを調べたところ，2 番目に近いものが「不快」である場合があった．このことから，誤認識した一因は最近傍決定則によって認識していることにあると考えられる．

表 3　連続音声の感情認識率

乳児音声区間検出 / 感情パターン	本法	手動
不快	3/10(30%)	5/10(50%)
空腹	13/14(92.9%)	10/15(66%)
眠気	2/5(40%)	5/5(100%)
全パターン	18/29(62.1%)	20/30(66.6%)

また，「眠気」の認識率が手動と比べて本法が低い要因の 1 つは，乳児音声区間の検出率が低いことにある．「眠気」を表す泣き声の場合，音声波形の振幅が小さい傾向があるので，振幅が閾値以下となり，乳児音声区間候補とならない可能性が他の感情より高い．

保護者による感情判定の精度が，本法での感情認識率の値の信頼性に影響を与えると考えられる．乳児が泣いた原因を，泣いた前後の状況で分類した研究[7]はあるが，著者らの知る限り，保護者による感情判定の精度を評価した研究は報告されていない．

5．まとめ

乳児音声に対し，Julius による単語信頼度と基本周波数の短時間での変化を特徴量とした乳児音声検出法を提案した．乳児音声の検出率は 6 割強となり，誤検出率は高々30%台にとどまった．また，本法により検出された乳児音声区間を用いて連続音声の感情認識を行った結果，62.1%という認識率を示した．また，手動で全音声区間を切り出し，感情認識を行ったものは，66.6%となった．このことから，乳児音声の感情認識において，本法が手動による乳児音声検出とほぼ同等の性能があったと考えられる．

乳児の泣き声には個人差があるので，本法の適用例を増やし，実用化への課題を整理する必要がある．本法では，Julius による単語認識率も 1 つの判定基準としているため，閾値を超える振幅をもつ大きな騒音を，乳児音声と誤検出する傾向がある．そのような騒音に再現性やパターンがあれば，ルールベースな方法を付加することで誤検出を減らすことが可能と考えられる．

参考文献

[1]http://www.mhlw.go.jp/topics/bukyoku/seisaku/syousika/index.html
[2]http://www8.cao.go.jp/shoushi/index.html
[3]菊池 健一郎, 荒川 薫, “低月齢乳児の泣き声周波数解析による啼泣原因推定～不快時の判定と眠気・空腹の考慮～”,信学技法 SIS2005-69, pp.55-60, 2006.
[4]三間 勇樹, 荒川 薫, “周波数解析による低月齢乳児の啼泣原因推定－空腹、眠気、不快の分類－”,信学技法 SIS2006-42, pp.43-46, 2006.
[5]西村 健吾, 三間 勇樹, 荒川 薫, “月齢一ヶ月乳児の泣き声解析による啼泣原因推定”,信学技法 SIS2006-87, pp.35-40, 2007.
[6]荒川 薫, 堀川 亮, 吉倉 忠, 伊藤 節子, “乳児の疼痛時における啼泣音声の特徴解析”, 信学技法 SIP2007-94, pp.65-69, 2007.
[7]三間 勇樹, 荒川 薫, “周波数解析による月齢二ヶ月乳児の啼泣原因推定”,電子情報通信学会総合大会講演論文集, pp. S-105-S-106, 2007.
[8]櫛田 康, 田伏 正佳, 吉冨 康成, “乳児の音声における感情的特徴のパターン認識”, ヒューマンインタフェースシンポジュウム 2008 論文集, pp.25-28, 2008.
[9]http://julius.sourceforge.jp/
[10] 林安 紀子, 出口 利定, 桐谷 滋, “選好振り向き法における 4～11 ケ月齢児の音声刺激に対する反応”, 音声言語医学, vol.37, no.3, pp.317-323, 1996.
[11] de Boysson-Bardies B vihman MM, “Adaptation to language: Evidence from babbling and first words in four languages; Language”, vol.67, pp.297-319, 1991.
[12] 天野 政昭, 近藤 公久, “親子の発声時間と発声速度の月齢変化”, 日本音響学会聴覚研究会資料, vol.32, H-2002-08, pp.57-62, 2002.
[13] 麦谷 綾子, “乳幼児の音声言語獲得”, 信学技報, vol.104, no.316, pp.13-18, 2004.
[14] レイ・D・ケント, チャールズ・リード, “音声の音響分析”, 海文堂出版, 1996.
[15] 中野 真里, 池添 史隆, 田伏 正佳, 吉冨 康成, “発声時の温度顔画像からの表情認識の効率化と性能への個人差の影響に関する検討”, 画像電子学会誌, vol. 38, no. 2, pp. 156-163, 2009.