感情音声コーパスの作成と評価 *

○ 飯田朱美[†], 明関賢太郎[†], Nick Campbell[‡], 安村通晃[†‡]
慶應義塾大学大学院 政策・メディア研究科[†‡]
(株)ATR 音声翻訳通信研究所[‡§]

1 はじめに

ATR で開発された自然音声波形接続型音声合成システム CHATR での利用を目的とした感情音声コーパスを作成し、評価を行なった。感情は人間のコミュニケーションにおいて重要な役割を担っており、感情表現が可能な合成音声が実現することで、話すことが困難な障碍者にとってのコミュニケーション・ツールへの応用も期待される。

本研究では、話者の自然な感情を引き出すために、喜び、怒り、悲しみの感情のこもった文章を朗読してもらい、収録した。特定の感情を一定時間持続させるため、対話形式ではなく、モノローグ形式の文章を収集した。

本稿では、感情音声コーパスの音響的特徴量の分析、及び、聴取実験の結果を報告する。また、作成したコーパスをデータベースとして、CHATR で合成した音声により、試験的に聴取実験を行なったので、その結果も併せて報告する。

2 感情を表現するテキストコーパスの作成

話者の喜び、怒り、悲しみの感情がこもった文章を新聞、WWW、障碍者による自費出版の日記などから収集し、CHATR のデータベースとして必要な文章量を収集した (表 1参照)。収集した文章は、読み手が感情を込めやすいようにそのままで用い、音素バランスは考慮しなかった。

表 1: 感情テキストコーパス

感情	記事数	文	モーラ数	音素数
喜び	12	461	21676	40916
怒り	15	495	21085	39171
悲しみ	9	426	16189	31840

テキストコーパスを構成している記事 (喜:12, 怒:15, 悲:9) を 2 部作成し、1 記事づつ、大学生 72 名に読んでもらい、感情を 3 択 (選択肢は喜, 怒, 悲) で判定してもらった。結果は 72 回答中、2 回答以外は収集者の感情分類と一致した (図 1)。

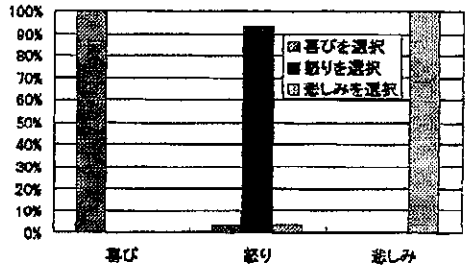

図 1: 感情テキストコーパスの評価

3 感情音声コーパスの音響的特徴

女性 1 名が無響室でテキストを朗読したものを収録し、16kHz、16bit でデジタル化した。表 2に感情別の基本周波数 (f0) の平均と標準偏差を示す。

表 2: f0 の平均と標準偏差

感情	平均	標準偏差
喜び	256.59	52.90
怒り	262.46	57.26
悲しみ	242.91	40.04

分析の結果、悲しみの平均 f0 は怒り、喜びに比べて低く、ダイナミックレンジも狭いことがわかった。次に、文中のポーズの継続時間長を測定し、3 感情の中で、悲しみのポーズが最も長いことが認められた (図 3)。これらは、短文を対象とする先行研究 [1] を支持する結果である (図 2)。

4 感情音声コーパスの評価

文脈の影響を極力抑えるために、感情音声コーパスを 1 文ごとにランダム提示し、大学生 29 名に 50 文ずつ感情を判別してもらった。喜び、怒り、悲しみの中から必ず 1 つ選択することとし、その上で、オプションで、次の項目をマークしてもらった (「よくわからない」「平

*Designing and testing a corpus of emotional speech
†Akemi IIDA, Kentaro MEISEKI, Nick CAMPBELL, Michiaki YASUMURA
‡Keio University
§ATR-ITL

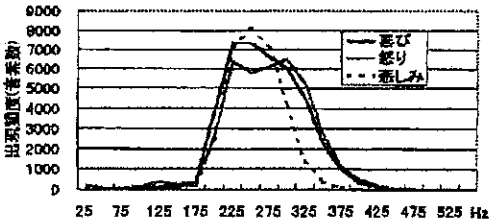

図 2: 感情音声別ピッチ分布

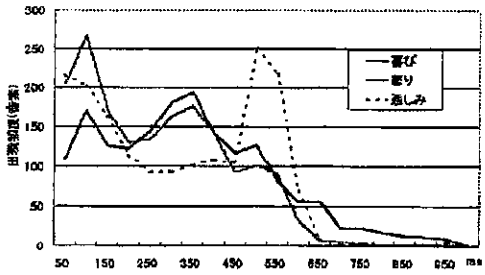

図 3: ポーズの継続時間長

坦」「文脈依存」「その感情に典型的な表現」)。これらは複数マークしても、全くしなくても良いこととした。感情音声コーパスは合計で 1382 文あり、全文につき 1 回答を得た (図 4)。結果は、喜び 80%、怒り 86%、悲しみ 93%の正解率で、1%の有意水準で検定を行なったところ、有意であった。

5 CHATR で作成した感情合成音声の聴取実験

本研究で作成した感情音声コーパスをデータベースとして、CHATR で、試験的に合成音声を作成し、聴取実験を行なった。各感情について、文脈にあまり依存しない内容の文を 5 文、大学生 18 名に感情判別してもらった。結果は、喜び 51%、怒り 60%、悲しみ 82%で、1%有意水準での判定結果は有意であった (図 5)。

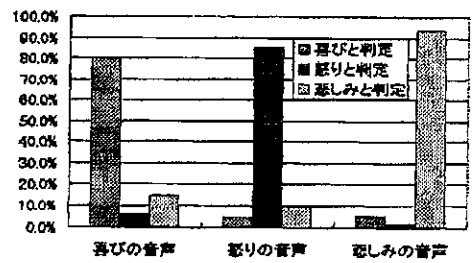

図 4: 感情音声の聴取実験結果

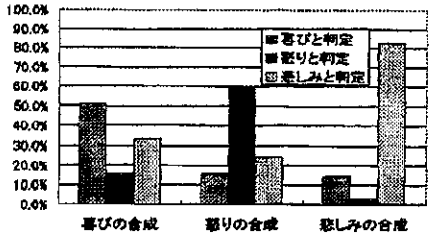

図 5: 感情合成音声の聴取実験結果

6 考察

感情音声コーパスの評価では、ランダムに提示したにも関わらず、被験者の 47%が、文脈の影響があると答えた。それに対して、CHATR の合成音声の場合は、13%であった。このことは、CHATR の合成音声の有意判定の結果が有意であることと合わせて考えると、被験者は文脈からではなく、音声情報から、合成音声の感情を判断したと考えられる。各感情に特有な音声情報をできるだけ多く引き出すために、文脈に依存した文章を収集して、合成音声のデータベースとする本研究のアプローチは有効だといえる。また、感情音声コーパスの評価では、回答の 23%について「平坦」がマークされており、CHATR の合成音声の評価でも、回答の 27% が「平坦」とマークされていた。しかし、これらの判定率もマークしなかったものと同様であった。このことは、音素そのものに、感情に関するかなりの情報が含まれていることを示唆していると考えられる。

7 まとめ

話者が感情を込めやすく、その感情を持続できるように配慮したテキストコーパスを作成し、話者の朗読を収録した。各感情音声コーパスのピッチのダイナミックレンジと、ポーズの継続時間長を測定したところ、先行研究を支持する結果を得た。感情音声コーパスの聴取実験の結果は有意であり、読み手の感情と聞き手が受ける読み手の感情とが一致した。合成音声の聴取実験の判定率は有意であり、被験者は音声情報から、合成音声の感情を判断したことがわかる。これにより、各々の感情特有の音声情報を多く含んだ感情音声コーパスの有効性を示しているといえる。今後は音素そのものに含まれる感情特性について注目していく予定である。

謝辞

ATR の芦村和幸研究員、太田洋子氏に感謝致します。また、音声処理をご指導頂いた東京大学広瀬啓吉教授と広瀬研究室の皆様に感謝いたします。

参考文献

[1] 北原義典,「音声における韻律の役割とその応用に関する研究」, 東京大学博士論文, 1996.12