

話者の感情印象の変化を抑制する話速変換手法*

江田 燎平[†], 小島 淳嗣[‡], 伊藤 克亘[†]

1 はじめに

従来の話速変換技術は、言語処理能力の衰えた高齢者や語学学習者の言語的理解補助を目的とした研究が多く、話速変換音声の明瞭度、話の聞き取りやすさなどを評価の基準として研究が行われてきた [1, 2, 3]。しかし、話速変換によって音声から伝達される情報のうちパラ言語情報の印象に変容を与えることが知られている [4, 5, 6]。これは、発話における言語情報と聞き手の感じた印象が一致しないことが起こると推測される。言語情報を正確に伝達することに重点をおいたニュースや報道番組に話速変換技術を用いた場合は、感情が変化した場合でも本来伝えたい内容を伝達出来るため問題とはならない。しかし、アニメーションやドラマなどの感情表現を多く含む番組を見る際には、その感情が変容することによって、本来伝えたい内容が伝達されないのが問題となる。

本研究では、アニメーションやドラマにおける話者の感情印象の変化を抑制する話速変換システムを作成する。

2 感情音声の収集

自然発話は、聴取者が話者の感情を主観的に評価することは困難であり、環境や精神状態に左右されやすい [7]。一方、アニメーションやドラマなどの演技音声では、主観的に感情を評価しやすい [8]。さらに、声優は声の演技方法に習熟しているため、環境や精神状態に影響されにくい。また、自然発話はその場にいる相手に発話を伝えることが目的であるため、聞き手は身振りや手振り、表情などでトータルで話者の感情状態を判断している。しかし、アニメーションの声優は相手が見えない中で声のみで伝えるため、身振り、手振り、表情は用いることはできない。さらに、感情表現するときには身振りや手振り、表情は演技者と絵で異なる場合がある。以上の観点から、自然発話に比べてアニメーションを扱う方が頑健性があると推測できるため、感情のラベル付け作業が容易である。よって、アニメーションを対象に音声データの収集を行う。

本研究では、日本アニメーションの「うさぎドロップ」を対象に感情音声を集める。「うさぎドロップ」は、1話につき約24分程度であり、全11話(約4.5時間)で構成されている。主に登場するキャラクターは2名(男1名,女1名)だが、その他に多くのキャラクターが存在するため、多くの感情音声を収集可能であると推測した。また、音声を収集するアニメーションの選定時に多くのアニメーションでは「悲しみ」の感情を含む音声が多く感じられた。つまり、「悲しみ」の感情音声を収集を十分に収集できない可能性がある。また、アニメーションの内容(構成)が「明るい」「楽しい」印象を得たものは、背景雑音が多く、評価実験で使用可能なサンプルの収集が困難であった。したがって、背景雑音の少なく、シリアスな内容構成となっている「うさぎドロップ」を対象に音声を収集することによって、音声の収集を行う。音声はサンプリング周波数44.1kHz、量子化ビット数16ビットでサンプルの抽出を行った。話速変換処理を行うときは16kHzにダウンサンプリングしている。また、本研究で収集する音声は複数の発話の重なりを持たないようにしている。アニメーション「うさぎドロップ」11話分から、知覚されたすべての感情喜び154発話、驚き53発話、焦り23発話、怒り110発話、悲しみ117発話、平静237発話、計694発話を収集した。

3 話速変換システム

文献 [6] では PICOLA を利用して感情保持を試みているが、PICOLA の性能は基本周波数 (F_0) の推定精度に依存するという問題がある。一方、WSOLA(Waveform Similarity

based OverLap-Add) [9] と呼ばれる SOLA(Synchronous OverLap-Add) をベースに改良をした話速変換手法では、相互相関関数により類似度の高い波形同士を足し合わせるため位相のずれが PICOLA に比べて少ないという利点がある。

そこで本研究では、WSOLA を改良して感情印象の変化を抑制する話速変換システムを作成する。

3.1 WSOLA アルゴリズム

WSOLA の手順について説明する。本研究ではフレーム幅 $L = 40[\text{ms}]$ 、シフト幅 $\text{shift} = 10[\text{ms}]$ として処理を行う。はじめに、伸縮パラメータ α を設定する。このパラメータは $\alpha = \text{shift}/H$ で表され、 H は波形セグメント A の中心時刻から波形セグメント B の中心時刻までの距離を表している。 $\alpha > 1$ であれば波形セグメント A' と波形セグメント B' が重ならないため、伸縮処理が行われる。一方、 $\alpha < 1$ であれば波形セグメント A' と波形セグメント B' が重なるため、波形の伸長処理が行われる。次に、入力された音声波形から、フレーム幅だけ切り出した波形を波形セグメント A とし、波形 A をシフト幅だけシフトさせた波形セグメントを A' とする。また、それぞれの中心時刻を S_{n-1} と $S_{n-1} + \text{shift}$ とする。波形セグメント A の中心時刻 S_{n-1} から、距離 H だけ離れた位置を中心とする波形セグメント B を波形セグメント B' とする。ここで、重畳加算時の位相のずれを考慮し、波形セグメント A' と類似度が最も高い波形セグメント B' の中心時刻を波形セグメント B の中心時刻 S_n からの距離 d を正規相互相関係数を用いて求める。次に、波形セグメント A の後半部分 ($S_{n-1} \sim S_{n-1} + \text{shift}$) と波形セグメント B' の前半部分 ($S_n - d \sim S_n$) に対し、それぞれ窓掛けを行う。窓掛け処理を行った後、この2つの波形セグメントに対して重畳加算を行う。最後に、波形セグメント B' を波形セグメント A として音声の終端までこの手順を繰り返す。

3.2 感情印象の抑制手法の導入

WSOLA に局所的に発話速度を制御する手法を導入し、感情印象の保持を行う。

感情の印象の知覚は基本周波数 (F_0) の変化の仕方に依存することが示されている [10]。また、発話におけるアクセントは F_0 の変化が大きく表れる箇所であり、感情の知覚に重要である箇所である。しかし、音声に対して話速変換処理を行うと局所的な F_0 の変化が変容するため、感情印象を保持するためにはアクセント箇所の F_0 の変化を一定に保つ必要がある。そこで本研究では、TILT [11] によってアクセント箇所の推定を行い、アクセントの箇所は発話速度を変化させず、その他の箇所は通常の処理を行うようにする。

3.2.1 TILT によるアクセント箇所の推定

本研究では、自己相関法を用いて F_0 を推定した。次に、推定した F_0 をメディアンフィルタで平滑化した後に無声区間に対して線形補間を行った。そして、求めた F_0 軌跡に対して TILT のパラメータを算出する。また、ここで求めた $Tilt_{dur}$ の値にはアクセント箇所のフレームインデックスが求まっている。そこで、このフレームインデックスを条件分岐を行い、範囲の中ならそのまま足し合わせ、外なら相互相関を用いて継ぎはぎ処理を行った。処理後の F_0 の軌跡は以下となった [図 1]。

4 話速変換音声の感情印象の評価実験

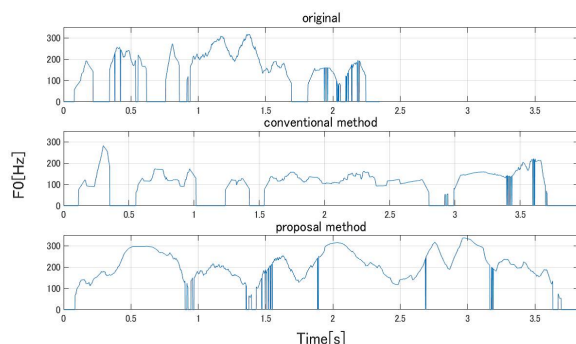
4.1 複数人でのラベル付け音声の収集

感情の知覚は人それぞれで個人差があり、1人でつけた感情ラベルの信頼性がない。また、評価実験の際に比較する感情音声を収集するため複数人での感情ラベルの付与を行う。

* Speech-Rate Conversion System to hold the Emotional Impression of the Speaker: Ryohei Edo (Hosei Univ.) et al.

[†] 法政大学 情報科学部

[‡] 法政大学 情報科学研究科

図 1. 話速変換処理前と後の F_0 の軌跡

アニメーション「うさぎドロップ」から収集した音声の中から各感情 15 発話、計 60 発話を対象に他 4 人の被験者を加えて感情ラベル付けを行い、1 人 (自分) のラベル付け結果との一致率を見た。

被験者は、聴覚機能の正常な大学生 5 名を対象に選び実験をおこなった。年齢は 21~22 歳である。被験者には音声の聴取する前、以下の注意喚起を行った。

1. 音声の途中で感情を判断するのではなく、最後まで聴取した後に感情ラベルの付与を行うこと。
2. 言語内容で感情を判断するのではなく、音声のみで感情を判断すること。

1 に関しては、「言語情報」で感情を判断せず、音声の感情状態のみで判断させるために、2 に関しては、発話の一部のみを聞いて感情を評価させないために注意喚起する。

音声の聴取は、密閉型ヘッドフォン (SONY MDR-ZX110-W) を利用し、聴取者の聞き取りやすい音量に各自で調整を行ってもらう。音声の再生は、PC 上の音楽ソフトウェアを利用し、聴取者自身の操作により次に再生する発話を聴取する。ただし、音声の聴取は 1 度のみとし、聞き直しは行わないようにした。また、発話内容のテキストは与えず、再生される音声のみで評価を行った。

収集した音声の中から各感情 15 発話、計 60 発話を対象に 5 人での感情ラベル付けを行い、最も多く知覚された感情ラベルを音声に付与した。その結果、怒り 16 発話、喜び 17 発話、悲しみ 15 発話、平静 15 発話、計 63 発話収集し、これを評価用音声とした。

4.2 感情印象の変化の評価実験

話速変換前と変換後で感情印象の変化を評価してもらい、感情印象が一致しているかどうか調査する。

被験者は、聴覚機能の正常な大学生 10 名を対象に実験を行った。年齢は 21~22 歳である。また、注意喚起および環境の設定は複数人でのラベルの付与の際に行ったものと同様である。複数人での感情ラベルの付与実験で評価した感情音声の中から「怒り」「喜び」「悲しみ」「平静」の 4 状態、各 5 発話、計 20 発話を選定する。これに対して伸縮パラメータ 0.6, 1.4 の 2 段階で話速変換処理を行い、計 40 発話を作成する。この 40 発話と収集した音声から適当に選定した 40 発話を加えて計 80 発話を評価用音声とする。

評価用音声から 1 発話ずつランダムに呈示させ、被験者に聴取させる。被験者には、それぞれが知覚した感情を「怒り」「喜び」「悲しみ」「平静」の中から 1 つ選び評価する。その後、話速変換前の音声 20 発話を聴取させ、上記と同様に評価してもらう。

従来手法と提案手法の結果はそれぞれ以下となった。

提案手法は従来手法と比べて 8.4% の感情印象の変化を抑制したといえる。これより、感情保持にはアクセント箇所の推定、および発話速度を局所的に変化させることが重要であると推測できる。

表 1. 従来手法の感情印象の一致率

	0.6 倍	1.4 倍	平均
怒り	76%	88%	82%
喜び	57%	61%	59%
悲しみ	83%	60%	69.5%
平静	58%	70%	64%
平均	68.5%	69.8%	69.1%

表 2. 提案手法の感情印象の一致率

	0.6 倍	1.4 倍	平均
怒り	78%	78%	78%
喜び	67%	78%	72.5%
悲しみ	93%	73%	83%
平静	79%	72%	75.5%
平均	79.5%	75.3%	77.5%

5 おわりに

今回、WSOLA にアクセント箇所の推定を導入し、話速変換による感情の変化を抑制するシステムの作成を行った。評価実験の結果、従来の話速変換システムより、提案手法は 8.4% 感情の変化を抑制することができた。したがって、提案手法は感情保持により最適な話速変換システムであるといえる。ただし、発話速度を局所的に制御したが、発話のモーラ数まで考慮に入れることが出来なかった。また、評価実験において聴覚機能の正常な大学生数名を対象に実験を行ったが、お年寄りなどの聴覚機能の衰えがある被験者を対象に行っていない。今後は、話者の感情印象の変化を抑制する話速変換システムに聞き取りやすさを考慮した話速変換システムを目指したい。

参考文献

- [1] 中村 章, 清山 信正, 池沢 龍, 都木 徹, 宮坂 榮一, “リアルタイム話速変換型受聴システム”, 音響誌, Vol.50, No.7, pp.509-520, 1994
- [2] 今井 篤, “ゆっくり話せば話がはい”, 信学技報, Vol.107, No.134, 85-89, 2007
- [3] 森田 直孝, 板倉文忠, “自己相関法による音声の時間軸での伸縮方式とその評価”, 信学技報, EA86-5, pp.9-16, 1986
- [4] 内田 照久, “音声の発話速度の制御がピッチ感及び話者の性格印象に与える影響” 音響誌, Vol.56, No.6, pp.396-405, 2000
- [5] 池原 仙一, 斎藤 博人, 武川 直樹, “発話速度変換会話音声の印象分析” 信学技報, HCS2012-1, HIP2012-1, pp.1-6, 2012
- [6] 小野 貴大, 斎藤 博人, 武川 直樹, “話者印象を保持するための韻律的特徴に基づいた区分的話速変換” 信学技報, HCS2014-7, HIP2014-7, pp.1-6, 2014
- [7] 森 大毅, 相澤 宏, 粕谷 英樹, “対話音声のバラ言語情報ラベリングの安定性”, 音響誌, 61(12), 690-697, 2005-12-01
- [8] 原 雄太郎, 伊藤 克亘, “時間構造分割特徴量に基づく感情発声の自動分類” 音響学研究発表会講演論文集, ROMBUNNO.1-7-4, 2011
- [9] W. Verhelst, M Roelands, “An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for high Time-Scale Modification of Speech”, Proc.IEEE, ICASSP-93, vol.II, pp.554-557, 1993
- [10] 高見 和之, 小林 聡, “2 モーラ語「ええ」に対するバラ言語情報の認知における F_0 の影響”, 信学技報, 108(116), 121-126, 2008
- [11] Taylor Paul, “Analysis and synthesis of intonation using the Tilt model”, J.Acoust.Soc.Am, 107(3), 1697-1714, 2000