F-029 波形重畳型音声合成における心理的態度の再現性評価

Reproducibility evaluation of mental attitude in concatenative speech synthesis

大久保 雅史 † 望月 亮 † ‡ 蓑輪 利光 ‡ 小林 哲則 †
Tadashi Okubo Ryo Mochizuki Toshimitsu Minowa Tetsunori Kobayashi

1. はじめに

PSOLA[1] 型の音声合成器において，発話者の心的態度を韻律の制御のみによってどの程度表現できるかについて検討した。

人間同士の対話において，発話には言語情報に加え心的態度に代表されるパラ言語情報が含まれ，それらが総合的に作用して発話者の意図を伝えている。高度な対話システムを実現するためには，音声合成器もまた言語情報のみならず発話者の心的態度を表現する必要がある。

合成音声によって心的態度を表現するためには，一般には分節的特徴と韻律特徴の双方を制御する必要があることが予想されるが，現在規則合成の代表的手法であるPSOLA 型の音声合成器においては，分節的特徴を変形することは難しく，可能な限りこれを抑えることが望まれる。効率的な規則合成器を作るためには，分節的特徴の制御が不可避となる条件について明らかにする必要がある。本研究ではこれを知るための手始めとして，原音声の分析の結果得られる韻律パターンを用いるとき，分節的特徴の変更を伴わないでどの程度原音の心的態度を再現できるかについて検討する。

2. 心的態度の表現

本研究では，心的態度の表現空間として，Russell[2] の提案する「快−不快」,「覚醒−不覚醒」の 2 軸を用いる。

従来音声合成における，パラ言語情報の付与に関しては，怒り，悲しみ，といった感情の扱いが中心であった[3][4]。しかしながら，実際の対話の場面で言語に加えて必要となる情報としては，相手発話に対する注目度・共感度，自発話に対する確信度，あるいは状態に対する満足度といった，対話状況に対する発話者の内的な評価であることが多く，それらは，基本感情とは必ずしも整合しない。一方，Russell の軸は，こういった内的評価と完全に一致するものではないものの，それぞれに関係を与えることができ，我々の目的には適している。

3. 評価実験

3.1 実験方法

発話内に４種類のフレーズを含む様々な態度を故意に含ませた音声を録音し、それぞれの音声に手動で音素ラベリングを行った。この音素ラベルに基づき音素時間長、音素内 4 点における基本周波数、パワーの韻律情報を抽出した。これらの自然な韻律情報を用いて、合成音声を作成した。合成音声作成の流れを図 1 に示す。

合成に使用した音声素片は、音韻バランスを考慮して約３時間の平静な読み上げによる音声コーパスを使用した。

†早稲田大学理工学部, Sch. of Sci. and Eng., Waseda Univ.
‡松下電器産業 (株), Matsushita Electric Ind. Co., Ltd

４種類の発話それぞれについて、原音声から１つ最も平静な状態の読み上げと思われる音声を選び、そのフレーズについての基準の音声とした。またすべての原音声、合成音声を評価対象音声とした。ランダムに評価対象音声を選択し、その発話の基準音声、評価対象音声という順に 9 人の評価者に聞かせた。

基準音声に対する評価対象音声の態度を、「快-不快」、「覚醒-不覚醒」の二軸において各軸それぞれ５段階で評価するよう指示した。

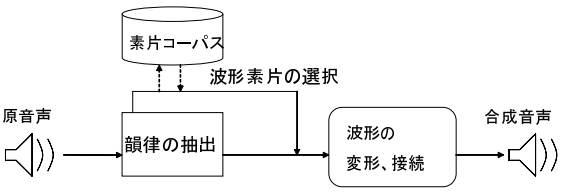


図 1: 合成音声作成の流れ

3.2 音声試料

発話内に自然に心的な態度が含まれるよう、応答の発話に決まったフレーズが含まれるような一対の対話文を複数作成した。フレーズは、対話の応答に現れるもの、フレーズの長さのバリエーションという観点から表 2 に示す 4 種類を選び、対話文は各フレーズにおいて、様々な態度が現れるような対話文を複数作成した。収録テキストの例を表 1 に示す。応答の発話の話者は固定として二人のプロの声優に読ませ録音した。

応答の発話の、決まったフレーズを含む部分を切り出して原音声試料とした。試聴実験に使用した音声試料を表 1 に示す。

表 1: 音声試料

発話内容	原音声の数
「自然のままの方がいい。」	14
「それは違うでしょ。」	14
「そうです。」	13
「つまりね。」	13

表 2: 収録テキストの例 (つまりね)

指定態度	発話内容
不快／覚醒	「ど、どういうことなの？(怒り)」 「つまりね、あんたがいけないの！」

3.3 実験結果

「快-不快」、「覚醒-不覚醒」それぞれに関して、各音声から合成音声を作成したときの、原音声の評価値の平均値に対する合成音声の評価値の平均値をそれぞれ図 2、図 3 に示す。また、図中の直線は、評価値の平均の一次近似直線を示す。

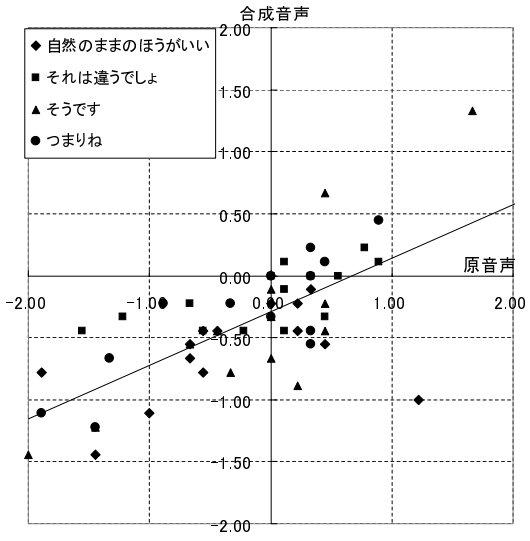


図 2: 快-不快に関する評価の変化

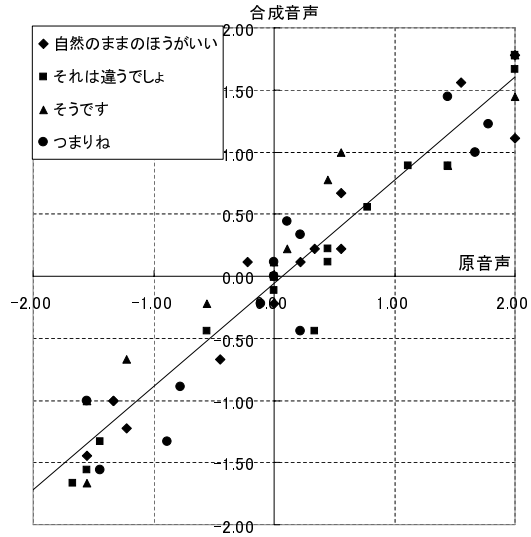


図 3: 覚醒-不覚醒に関する評価の変化

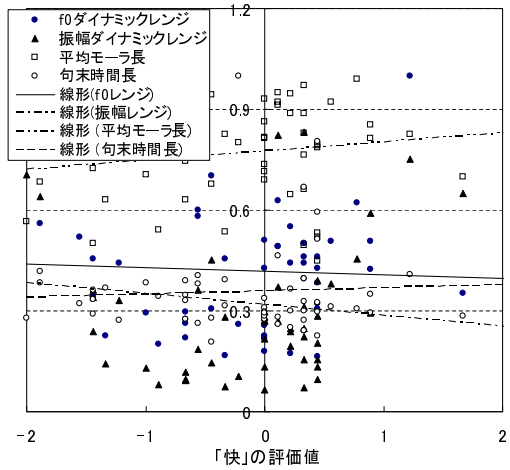
実験結果より、発話文の違いや、基準音声の違いによる評価への影響は多少あるものの、全体的にほぼ同じような傾向を示したと言える。原音声での「覚醒-不覚醒」の評価値は、その韻律で作成した合成音声においてもよく再現されているが「快-不快」の評価値は合成音声においては弱まっている。

図 4 に原音声の快、覚醒の評価値と韻律パラメータとの関係を示す。これらより、「快-不快」の評価値には「覚醒-不覚醒」ほどの韻律の影響はないものと考えられる。

これらより、「快-不快」の合成音声は、韻律情報のみから作成することが難しいと言える。

4. むすび

心的態度を表現できるような高度な対話システムを作ることを目的とし、PSOLA 型の音声合成器において、韻律情報のみから心的態度をどの程度表現可能か検討を行った。心的態度の表現尺度として、「快-不快」「覚醒-不覚醒」の二軸を用いた。



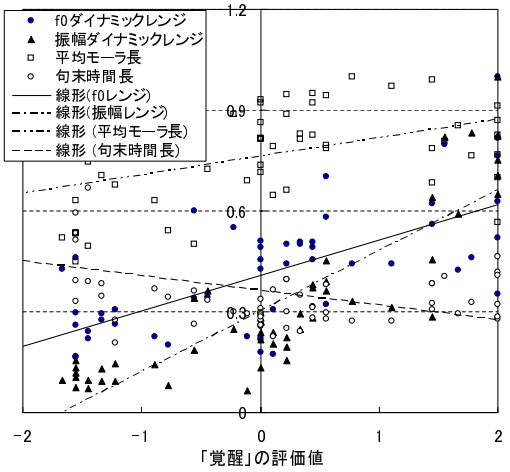


図 4: 各評価値と韻律パラメータとの関係

様々な態度を含む音声から韻律を抽出して合成音声を作成し、原音声と合成音声を試聴実験により二次元平面上に評価させ、どのような態度がどの程度合成音声において再現出来るか試聴実験を行った。

結果として合成音声においては、「覚醒-不覚醒」については十分表現可能だが、「快-不快」については表現が薄らいでしまうというデータが得られた。これより、「快-不快」の表現のためには、それぞれの態度に特化した素片の作成が必要であると考えられる。今後それらの素片を使用した上での実験も行う必要がある。

参考文献

[1] Charpentier F., Moulines E. “Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, Pros. Eurospeech ’89,1989.

[2] Russell J.A. “A circumplex model of affect.”, Journal of Personality and Social Psychology,39,1161-1178,1980.

[3] Y.Kitahara,Y.Tohkura “Prosodic Control to Express Emotions for Man-Machine Speech Interaction.”,IEICE Trans. Fundamentals,E75-A,No.2,155-163,1992.2.

[4] Iida A. Campbell N. Iga S. Higuchi F. and Yasmura M. “A Speech Synthesis System with emotion with assisting communication.”, In Proceeding of the ISCA Workgroup on Speech and Emotion,167-172,2000.