Vol.2015-MUS-107 No.70
2015/5/24

# 感情音声変換を目的とした Valence-Activation 2 次元感情空間と関連する音響特徴量推定と基本周波数制御法の検討

濱田康弘[†1]　Elbarougy Reda[†2]　赤木正人[†3]

概要：本発表では、Valence-Activation 2 次元感情空間 (V-A 空間)上での感情音声変換を目的とし、関連する音響特徴量推定法と基本周波数の制御法について報告する。 音響特徴量推定では Fuzzy Inference System を用いて V-A 空間と音響特徴量（基本周波数の平均値、最大値、上昇時の傾き）を関連付け、藤崎モデルを用いて基本周波数の制御を行う。これにより、推定された音響特徴量を満足する基本周波数軌跡が表現された。

## 1. はじめに

Story teller system や speech-to-speech translation system (S2ST) 等の，人が発話した音声をコンピュータが認識し，コンピュータが内容を理解し音声で応答するシステムは，人とコンピュータを繋ぐマンマシンインターフェイスとして重要である。これを実現する為に，これまでに様々なアプローチが行われてきているが，言語情報以外の情報であるパラ・非言語情報 (感情，個人性，性別等) [1] に関しては未解決の部分が多い。人が話すような自然なシステムを構築する為には，パラ・非言語情報の一つとして感情を考慮したシステム (Affective S2ST) [2] を構築することが必要である。

本稿では，パラ・非言語情報の中でも特に感情に焦点を当て，平静音声から感情情報が付与された音声へと変換する方法について検討する。

多くの研究において感情は｛喜び，怒り，悲しみ｝等のカテゴリカルなものとして捉えられている。しかしながら，人の多様で複雑な感情をシステムに組み込む為には単一のカテゴリに分類されるものではなく，多次元空間で連続的に変化するような感情を表現することが望ましいと考えられる。心理学等の研究から，Valence (快-不快)，Activation (活性-非活性) の 2 軸は感情を表現する主要な軸であることが報告されている [3]。Valence-Activation 2 次元感情空間 (以下，V-A 空間と呼ぶ) において，例えば｛喜び｝の感情は図 1 のように Valence, Activation と共に正の値をとる第一象限に布置され，この中で喜びの度合いが連続して変化している。このような多様で複雑な感情を表現する為に，本研究では感情を V-A 空間上の連続空間として捉える次元的アプローチをとる。次元的アプローチにより，Elbarougy らは音声の感情認識を行い，精度の高いシステムを構築している [4]。このことから次元的アプローチは有効であり，感情音声変換にも適用できると考えられる。

方法として，V-A 空間と音響特徴量との関係を Fuzzy Inference System (FIS) [5] によって構築し，V-A 空間の布置情報を入力として FIS により推定された音響特徴量を用いて，平静音声を変形する。本稿ではまず初めに FIS を用いた音響特徴量推定について述べる。次に推定された音響特徴量をターゲットとした基本周波数の制御法について検討し，最後にまとめを行う。

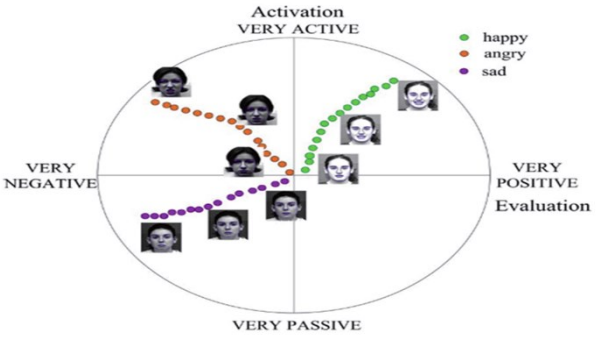


図 1. Emotion space spanned by Valence-Activation axes

## 2. 音響特徴量推定法

### 2.1 音声データ

音声データは｛平静，喜び，悲しみ，抑えた怒り，荒げた怒り｝含んだ 1 話者の文章発話 (計 179 個) である富士通データベースを用いた。

### 2.2 音響特徴

本研究では 21 種類の音響特徴が用いられた。それぞれ，*F0* に関連した特徴 (4 種類)，パワー包絡に関連した特徴 (4 種類)，スペクトルに関連した特徴 (5 種類)，継続時間長に関連した特徴 (3 種類)，音質に関連した特徴 (5 種類) である。これらは Elbarougy らが用いたもの [4] と同様である。本稿では特に *F0* に関連した特徴である，*F0* の平均，最大値，フレーズ毎の正の勾配，第一フレーズの勾配と継続時間長に関連した特徴である，全体の時間長，子音の長さ，子音と母音の長さの比について考慮する。

†1 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

†2 Damietta University

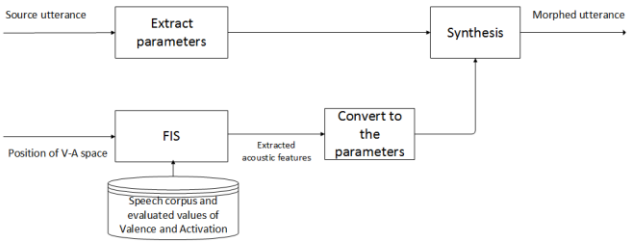


図 2. Proposed emotional speech conversion process

### 2.3 FIS を用いた音響特徴量推定

V-A 空間と音響特徴量の関係を記述するために, Adaptive network Fuzzy Inference System (ANFIS) を用いた。ANFIS により，V-A 空間上の布置情報から音響特徴量を推定することが可能となる。富士通データベースの 179 個の各音声について 21 種類の音響特徴を抽出した。音響特徴の抽出には分析合成系 STRAIGHT [6] を用いた。抽出した音響特徴量と聴取実験によって得られた Valence, Activation の評価値を ANFIS により学習した。ここで，179 個の音声データの 80%を学習データ，20%をテストデータとして用いた。

## 3. *F0* に関連する特徴の制御法

*F0* に関連する各特徴はそれぞれ独立ではない為，各特徴量を満足する *F0* 軌跡へと変形する制御規則が必要である。図 2 に提案する感情音声変換システムの概略を示す。*F0* に関連する特徴の制御の方略として，V-A 空間の布置情報から推定された *F0* に関連する特徴量を元に，藤崎モデルを用いてパラメータ変換することで，ソース音声の *F0* 軌跡の変形を行う。

### 3.1 藤崎モデル

藤崎モデル [7] は音声の生成メカニズムに基づいた F0 パターンをフレーズ成分，アクセント成分，ベースラインの和で表現した数学的モデルである。*F0* の対数値は以下の式で記述される。

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i G P_i(t - T_{0i}) + \sum_{j=1}^{J} Aa_j \{ Ga_j(t - T_{1j}) - G_{aj}(t - T_{2j}) \} \quad (1)$$

*Gp* はフレーズ成分に相当するインパルス応答，*Ga* はアクセント成分に相当するステップ応答であり，以下の式によって記述される。

$$G_{pi}(t) = \begin{cases} \alpha_i^2 \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2)$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma\right], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (3)$$

*Fb* は基本周波数のベースラインであり，*Api* は *i* 番目のフレーズ指令の大きさ，*T0i* はその起点，*Aaj* は *j* 番目のアクセント指令の大きさ，*T1j*, *T2j* はその立ち上がりと立下り時点を表わしている。$\alpha$，$\beta$は臨界制動系の角周波数であり，凡そ$\alpha = 1.0/s$，$\beta = 20/s$で表される。本稿では Mixdorff によって提案された方法 [8] を用いて藤崎モデルのパラメータを抽出する。

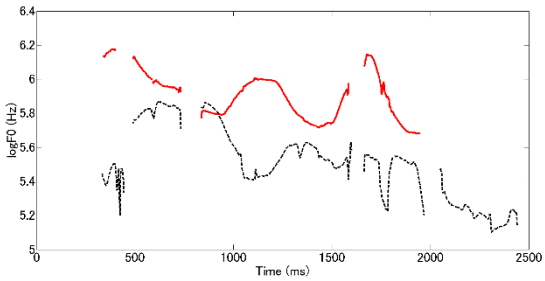


図 3. F0 tragectory of a source speech (dashed) and synthesized speech (solid)

### 3.2 藤崎モデルによる *F0* 軌跡の制御

本稿では FIS から推定された *F0* に関連する特徴 *(F0* の平均，最大値，勾配) を藤崎モデルのパラメータ値に変換し，ソース音声の *F0* 軌跡をターゲットとする *F0* 軌跡に変形する。

まず始めに FIS により推定された時間長に関連する特徴 (全体の長さ，子音の長さ，子音と母音の比) を元に音声を伸縮する。次にソース音声から得られた時間軸に関連する藤崎モデルのパラメータ *T0*, *T1*, *T2* を時間長の伸縮にあわせて変換する。*Fb*, *Ap*, *Aa* はターゲット音声の値と変換することによって決定する。図 4 に藤崎モデルのパラメータにより制御された *F0* 軌跡を示す。

ソース音声，ターゲット音声，再合成された音声の *F0* の平均，最大値を算出した結果，対数スケールでの *F0* の平均は，ソース音声で 2.39，ターゲット音声で 2.56，合成音声で 2.55 であった。また *F0* の最大値は，ソース音声で 2.55，ターゲット音声で 2.70，合成音声で 2.68 であった。この結果から，合成音声の *F0* の音響特徴はターゲット音声の音響特徴に近いことが明らかとなった。よって，藤崎モデルのパラメータを制御することにより，感情音声の *F0* 軌跡を表現出来ることが示された。

## 4. おわりに

本稿では V-A 空間上での感情音声変換を目的とし、関連する音響特徴量推定法と基本周波数の制御法について報告した。 音響特徴量推定では ANFIS を用いて V-A 空間と音 *F0* の平均値、最大値、上昇時の傾きを関連付け、藤崎モデルを用いて基本周波数の制御を行った。その結果，藤崎モデルにより推定された音響特徴量を満足する基本周波数軌跡が表現出来ることが示された。

**謝辞** 本研究の一部は，科研費 (25240026) および A3 Foresight Program (JSPS) の援助を受けて行われた。

## 参考文献

1) H. Fujisaki, “Information, Prosody, and Modeling - with Emphasis on Tonal Features of Speech – ,” Proc. Speech Prosody 2004,

pp. 23-26, 2004
2) .M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li “Toward Affective Speech-to-Speech Translation: Strategy for Emotional Speech Recognition and Synthesis in Multiple Languages ,” Proc. APSIPA 2014,
3) J. A. Russell and P. Geraldine, “A Description of the Affective Quality Attributed to Environments ,” J. Pers. Soc. Psychol., 38 (2), 311-322, 1980.
4) R. Elbarougy and M. Akagi, “Improving Speech Emotion Dimensions Estimation Using a Three- Layer Model for Human Perception ,” Acoust. Sci. Tech., 35 (2), 86-98, 2014.
5) J.-S.R. Jang, “ANFIS: Adaptive-Network-Based Fuzzy Inference System ,” IEEE Trans. Syst., Man. Cyberm., 23 (3), 665-685, 1993.
6) H. Kawahara et al :, “Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds ,” Speech Commun., 27, 187-207, 1999.
7) H. Fujisaki, “Manifestation of Linguistic, Paralinguistic and Non-linguistic Information in the Prosodic Characteristics of Speech ,” Tech. Rep. of IEICE, 37 (9), pp.1-8, 1994.
8) H. Mixdorff. “A Novel Approach to The Fully Automatic Extraction of FUJISAKI Model Parameters,” Proc. ICA 2000, 3, pp. 1281-1284, 2000.