実世界情報システムプロジェクト
視聴覚研究グループ峯松研究室 聴知覚的インタフェース研究ユニット

峯松 信明

情報理工学系研究科電子情報学専攻

概 要

本研究ユニットは，人間と計算機間の意志疎通を「音声対話」に基づいて行なうための技術構築を行なっている。特に「話し言葉」が持つ「書き言葉」に無い特性に着目し，ユーザ発話からのパラ言語・非言語情報の抽出，計算機発話に対するパラ言語・非言語情報の付与を検討している。具体的には，音声音響情報からの知覚的年齢の推定，感情が付与された合成音声の生成，その場の状況を考慮した対話管理，などについて検討している。

1 はじめに

人間と計算機（ロボット）が共存する社会。SF映画の中では常識となっている未来像である。そこでは人と計算機の間に主従関係こそあれ，人が計算機の中に「人格」を感じる世界が描かれている。人は何をもってオブジェクトの中に「人格」を感じるのか？筆者はそれは「良い意味での裏切り」であると考えている。相手の行動が100%予測できる場合，その行動が如何に精巧であれ，それは単なる「プログラム」であって，そこに「人格」を感じることはない。相手の行動が予測できなかった場合（裏切られた場合），そして，その行動をとった相手の「意図，真意」を事後的に理解できた場合，人はその行動に相手の「人格」を感じ,「気の利いた奴」として認めるようになる。

日本は，地理・地形的に他国（他文化）との交流が遮られ，また歴史的にその交流を遮ってきた経緯を持つが，その結果日本人は，世界的に見ると非常にユニークなコミュニケーション手段を，当たり前のように行使する民族となっている。「阿吽の呼吸」という言葉で表される，その場の「雰囲気」，相手の発言の裏に隠された「真意」を瞬時に察し，それに応じた行動をとる能力が尊ばれ，それが出来て初めて「気の利いた奴」となる。

「気の利いた奴」を対話システム上に実現する場合，ユーザが明示的に “言葉として”発した意図のみならず，非言語的な情報を汲み取る必要がある。と同時に，ユーザに明示的に “言葉として”返すのではなく，非言語的なメディアを使って間接的に提示する能力が要求される。本研究では，前者に対してはユーザの年齢を音声情報から知覚する枠組みについて検討し，後者に対しては合成音声に感情を付与する枠組みについて検討した。

2 音声からの知覚的年齢の推定

年齢の自動推定を考える場合，話し手の実年齢と，その音声を聴取した時に聞き手が感じる年齢が考えられる。本研究では後者，即ち知覚的年齢を対象とする。これは，聞き手が相手の年齢を察知する場合，後者がその対象となるからである。

2.1 聴取実験による知覚的年齢の推定

年齢幅の広い（6〜90歳）音声DBに対して，話者毎の知覚的年齢ラベリングを行なった。被験者は20歳前後の成人30名である。なお，男声のみを使用した。被験者は各話者の音声を一文聞き，その年齢を1歳単位で推定する。合計約500名

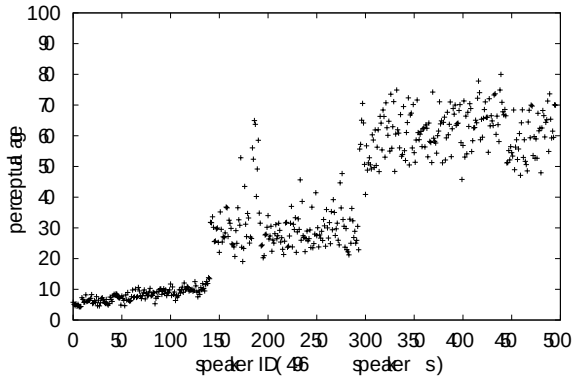

図 1: DB 話者毎の平均知覚的年齢

の話者の年齢を推定させた。各 DB 話者に対する（被験者間の）平均知覚的年齢を図 1 に示す。

2.2 話者モデルを用いた知覚的年齢の推定

本研究は，前節で行なった人間による知覚的年齢推定プロセスのシミュレーションである。これを話者認識技術を用いて実装する。話者モデルであるが広く一般的に使われる GMM モデルを採用した。約 500 名の DB 話者各々に対して，約 60 秒の音声から（無音区間は排除）GMM を作成した。

以下の式で知覚的年齢（PA）推定を検討した。

$$PA = \frac{\sum P(x|o) \times x}{\sum P(x|o)} \tag{1}$$

ここで o は音響観測量，x が（聴取実験より定義される）知覚的年齢である。即ち o に対する事後確率を用いた期待値操作で知覚的年齢を推定する。ここで $P(x|o)$ はベイズ則により

$$P(x|o) = \frac{P(o|x)P(x)}{P(o)} \tag{2}$$

と変形され，$P(o)$ を定数項と考えれば期待値操作の重みは $P(o|x)P(x)$ となる。$P(x)$ は知覚的年齢に対する事前確率であり，現時点でそれを仮定することは不可能であり，本研究では $P(x)$ として一様な分布をもつ確率密度関数を考える。以上の結果，$P(o|x)$ を重みとした期待値操作へと帰着される。但し，本研究で利用する DB 話者の知覚的年齢分布は明らかに偏りがあり，それを是正する処理を必要に応じて導入することになる。

2.2.1 DB 話者に年齢ラベルを与える場合

聴取実験の結果から，各 DB 話者毎に知覚的年齢値をラベルとして付与することが可能である。その情報から $P(o|x)$ を推定する。まず DB 話者を平均知覚的年齢毎に分類し，各年齢毎に尤度 $P(o|x)$ を以下の方法で近似した。知覚的年齢が x 歳である i 番目の話者のモデルを $M_x(i)$ とする。

$$P(o|x) \quad = \quad \max_i P(o|M_x(i)) \tag{3}$$

しかし年齢幅は有限であるため，最終的に得られる期待値の（年齢幅中の）位置によっては，期待値前後の年代幅のレンジに偏りが生じ，最終的な推定結果に偏りが生じる。そこで，上位 $P(o|x)$ の N 年齢についてのみ期待値操作を行なった。

2.2.2 DB 話者に年齢分布を与える場合

聴取実験の結果から，各 DB 話者に対して知覚的年齢の分布を付与することが可能である。この場合，分布が広がっていればその話者は「年齢不詳」度が高い，ということになる。なお，知覚的年齢に対する DB 話者数分布の偏りを解消するため，ここでは，入力話者が全 DB 話者と等しい距離に置かれている場合，知覚的年齢分布が一様となるよう，正規化関数を導入した。

ラベル及び分布として年齢分布を与えた場合の推定結果を図 2 に示す（上：ラベル，下：分布)。分布として与えることにより，推定精度が上がっていることが分るが，推定が大きくずれる話者がいることも分る。先行研究より，話者モデリング技術のみでは技術的限界が示唆されており，話速やパワーの局所的変動など，韻律的特徴も考慮した話者モデリングが必要であろう。

3 韻律的制御に基づく感情の生成

本研究室では従来より，読み上げ調の音声合成に対して，より自然な韻律付与を行なうべく，基本周波数（F_0）パターンのモデル化及び，テキストからのモデルパラメータの自動推定に対して研

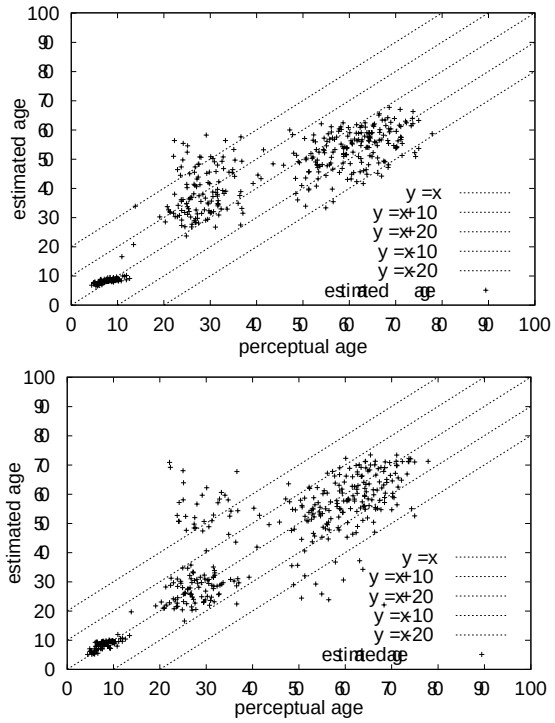

図 2: 知覚的年齢推定結果

究を行なってきた。本研究では，その対象を感情音声にまで広げて実験的検討を行なった。

3.1 F_0 パターン生成モデル

音声波形から観察される F_0 パターンを対数軸で見ると，なだらかな起伏の上に局所的な起伏が乗った形として捉えることができる。このなだらかな起伏は句頭から句末に緩やかに下降するフレーズ成分に対応し，局所的な起伏は語句のアクセント型を表現するアクセント成分に対応するとして，これらの成分が式 (4)，(5) のように離散的な指令に対する臨界制動 2 次線形系応答として生成され，式 (6) のように重畳したものであると考えるのが，F_0 モデルである（図 3）。

$$G_{p_i} = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (4)$$

$$G_{a_i} = \begin{cases} \min[1-(1+\beta_j t)\exp(-\beta_j t), \theta] & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (5)$$

$$\begin{aligned} \ln F_0(t) &= \ln F_b + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0_i}) \\ &\quad + \sum_{j=1}^{J} A_{a_j} \{G_{a_j}(t - T_{1_j}) - G_{a_j}(t - T_{2_j})\} \end{aligned} \quad (6)$$

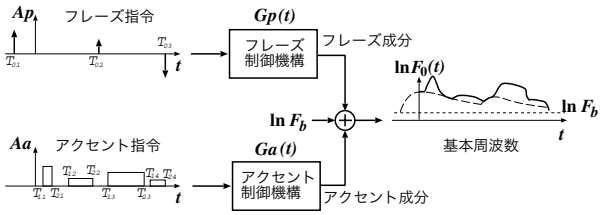

図 3: F_0 パターン生成モデル

本モデルは，少ないパラメータで基本周波数パターンを良く近似し，パラメータと言語情報との対応が良いので，TTS システムに用いることができるなどの利点が多いが，音声波形からのパラメータ自動抽出が難しいという欠点もある。

3.2 テキストからの感情音声用 F_0 モデルパラメータの予測とその精度

テキスト音声合成に本モデルを用いる場合，与えられた漢字仮名混じり文から言語情報，音韻情報を抽出し，それに基づいて本モデルパラメータの値を推定し，それに従って最終的な F_0 パターンを描くことになる。紙面の関係上，詳細な説明は省くが，対象とする部位の文中の位置，アクセント句中の位置，品詞，活用形，アクセント型などの言語情報を下にフレーズ指令，アクセント指令に関する各種パラメータを推定することになる。予測のための枠組みとしては，正解が与えられたデータベースに対して，決定木あるいは回帰木を用いて対象となるパラメータ値を推定した。

表 1 に推定パラメータと正解パラメータ間の平均誤差の様子を，また，表 2 に最終的に得られた推定 F_0 パターンと実際に各種感情で発声された同一文の F_0 パターンとの平均誤差について示す。なお，用いた音声資料は女性 1 名が各種感情で読み上げた疑似感情音声である。これらの結果より，平静の読み上げ音声と変らぬ精度で感情音声に対する F_0 パターンが生成可能であることが分る。

4 「場」の理解に基づく対話管理

従来の音声対話システム，あるいは音声インターフェースは，大規模 DB に対して情報検索

表 1: 推定されたモデルパラメータの平均誤差

		平静	怒	喜	哀
A_p	close	0.20	0.21	0.22	0.20
	open	0.21	0.22	0.22	0.20
T_{0off}	close	0.11	0.18	0.13	0.13
	open	0.11	0.20	0.13	0.13
A_a	close	0.17	0.16	0.17	0.12
	open	0.16	0.16	0.16	0.12
T_{1off}	close	0.09	0.09	0.10	0.09
	open	0.09	0.09	0.09	0.10
T_{2off}	close	0.06	0.06	0.06	0.07
	open	0.07	0.06	0.06	0.07

表 2: 各感情における推定 F_0 パターン誤差

	close	open
平静	0.041	0.049
怒	0.054	0.049
喜	0.068	0.050
哀	0.053	0.048

を行なう時の入出力メディアの一つとして捉えられることが多かった。この場合，システムが「知らない」ことは，DB（即ち世界知識）の更新が行なわれない限り未来永劫「知らない」ままである。しかし人と人との対話は，互いが不完全な知識しか持ち合わせていない場合でも，それを補完し，対話を通して（即席ではあるが）世界知識を構築し，それを前提として問題解決を図ることが多い。本研究では，そのような仕組みを対話管理部に持たせ，特に応答音声においてどのような韻律的脚色を施すことが世界知識の効率的構築に貢献するのか，について実験的に検討した。

5　その他の研究成果

人と人との対話には様々な側面が存在する。上述した研究以外にも本テーマに関係する研究成果が出ており，以下に簡単に説明する。

感情音声の認識　機械との対話がより人間味を帯びてくると，当然そこには感情が生まれる。従来の音声認識は読み上げ音声を対象としており，感情音声は未知なる分野である。ここでは，話者適応技術を用いて各種感情に対して音響モデルを適応することで認識率の向上を実現した。

話題追従型の言語モデル　対話が弾むと感情が生まれ，その内容も多岐の渡るようになる。このような音声を認識する場合，静的な言語モデルでは十分な言語制約が生成できない。ここでは，話題追従型の言語モデル適応を検討しており，疑似的に生成した「支離滅裂」文セットに対してより高精度に言語制約を生成できることを確認した。

F_0 モデルパラメータ推定　韻律生成時に必要となる F_0 モデルであるが，与えられた音声に対するパラメータ自動抽出精度は十分に高いとは言えない。特に感情音声の場合，精度劣化が激しい。ここでは特に，色のついた音声を対象にしてパラメータ抽出の高精度化を検討している。

6　まとめ

次世代の音声対話インターフェースには必須の要素である「パラ言語，非言語情報に対する処理系」に焦点を当て，各種検討を行なった。今後，各技術精度の向上と共に，新たな情報源，メディアに対しても積極的に導入を検討していきたい。

発表文献（一部）

[1] 峯松信明，広瀬啓吉，関口真理子，“話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定”，情報処理学会論文誌，vol.43, no.7, pp.2186–2196 (2002).

[2] A. Sakurai, K. Hirose, and N. Minematsu, “Data-driven generation of F_0 contours using a superpositional model,” Speech Communication (2003, to be published).

[3] N. Minematsu, M. Sekiguchi, and K. Hirose, “Automatic estimation of one’s age with his/her speech based upon acoustic modeling techniques of speakers,” Proc. ICASSP, vol.1, pp.137–140 (2002).

[4] 山内景太，峯松信明，広瀬啓吉，“話者認識技術を応用した知覚的年齢分布の自動推定”，電子情報通信学会音声研究会，SP2002-186，pp.43–48 (2003)

[5] 桂聡哉，広瀬啓吉，峯松信明，“感情音声合成のための生成過程モデルに基づくコーパスベース韻律生成とその評価”，電子情報通信学会音声研究会，SP2002-184, pp.31–36 (2003)

[6] 多胡順司，広瀬啓吉，峯松信明，“エージェント対話システムのための対話処理と応答文生成”，情報処理学会春季全国大会，5T7B-4 (2003，発表予定)