音声情報処理におけるパラ・非言語情報*

○ 広瀬 啓吉　（東京大・新領域）

１．はじめに

近年のマンマシンインターフェースの進展に伴い、研究者の関心が、それまでの論理情報から、感性情報の表現・伝達にも向けられるようになってきた。わが国では、1992 年度から 3 年間に渡って行われた文部省重点領域研究「感性情報処理の情報学・心理学的研究」で、感性が工学的な見地から取り上げられ、顔の表情、音声等と感情等との関連が定量的に調べられた。音声に関してみれば、その後、感情音声の分析、合成、認識の研究が活発に行なわれるようになっている。

意図・態度、喜び・悲しみといったパラ言語・非言語情報は、言語能力が未完な幼児でも表出しているものであり、言語情報と比べ先天的なものということができる。人間間のコミュニケーションではごく普通に表出・伝達が行なわれているが、これを機械で実現することは言語情報の場合と比べ必ずしも容易でない。これは、論理情報である言語情報は定式化になじんだものであるのに対し、感性情報であるパラ言語・非言語情報はその定義、分類にすらあいまいな部分があるためである。

ここでは、主として感情に焦点を当て、まずパラ言語・非言語情報についてその音響的特徴を簡単に整理した後、音声合成あるいは認識でどのように取り扱われているかについて概説するとともに、今後の研究について考察する。

２．特徴の分析

2.1 感情とは

工学的な観点から感情を扱う場合、喜び、怒り、悲しみなどの各種の感情を区別して取り扱うことになる。しかしながら、実際に感情の種類がどのように定義されるかについては必ずしも明確になっていない。これは、感性的な情報を、論理的な言語情報に対応付けることの難しさを表わすもので、普遍的で明快な対応付けというものは存在しないといった方が正確であろう。一般には、独立性が高く、顕著に表出される基本感情として、恐れ、怒り、喜び、悲しみ、驚き、嫌悪があるとされ、これらを対象とした研究が多く行なわれているが、特に怒りについては激しい怒り（hot anger）と押し殺した怒り（cold anger）を、音響的な面からも別のものとして取り扱う必要があるのは周知の事実である。この他、軽蔑なども基本感情として捉えられている。一方、感情を表わす 142 もの用語が定義されているが、それらのほとんどは、互いに関連性の高い２次的な分類となっているため、扱いにくいという側面がある。ただし、注意しなければならないのは、このような 2 次的に分類された感情が、実際の応用では重要になってくるということがある[1]。喜びを取り扱う対話音声システムというよりは、楽しそうな会話を実現する対話システムという方が自然でもあり、実際、必要とされてもいるであろう。

2.2 音響的特徴

パラ言語・非言語情報は一般に韻律との関係が深い。表 1 は、韻律が、言語情報、パラ言語・非言語情報の伝達で果たす役割の一覧である。言語情報の伝達では分節的特徴の補助的な役割であった韻律的特徴が、パラ言語・非言語情報の伝達では主要な役割を果たしている。

Table 1.　Information transmitted by prosodic features.

Information	Content
Linguistic	Word meaning 　Stress, Accent type, Tone Syntactic structure Interrogation Emphasis Paragraphing
Para-linguistic	Speaker's attitude 　Polite, Rude, etc. Speaker's intention 　Inquiry, Decisiveness, etc.
Non-linguistic	Speaker's emotion 　Anger, Joy, etc. Individuality Dialogue act 　Filled pause, Restatement, etc.

ここで、韻律を表現する音響的特徴について明確にしておきたい。韻律的特徴は主に音源に関連した特徴で、音源振動の基本周波数とパワー、発話速度、母音の韻質等とされる。ここで、実際の音響的な特徴量として、パワーとしては波形の短時間パワー、発話速度としては音素の持続時間、母音の韻質としてはフォルマント周波数等が用いられることが多いが、これらが直接、韻律を表現するものではなく、標準的な値からの偏差が韻律情報を伝達することに留意する必要があ

* Para- and non-linguistic information in speech information processing.
By Keikichi Hirose (Graduate School of Frontier Sciences, University of Tokyo)

る。これに対し、基本周波数は、それが直接的に韻律を表わす特徴となる。また、韻律的特徴は、アクセント型や構文といった基本的な言語情報の表現に用いられ、それによる制約を受ける。その制約下でパラ言語・非言語情報に対応した特徴の変形が行なわれる。このため、感情音声の特徴を論じるとき、平静音声のそれとの比較が必要になる。

パラ言語・非言語情報と韻律的特徴との関連は早くから調べられ、例えば、文単位の基本周波数や発話速度が着目されてきた。例えば、喜びでは基本周波数が高く、悲しみでは低くなることはよく知られている。あるいは、基本周波数についてはその時間的な変化の特徴も論じられている[2]。しかしながら、これだけでは韻律的特徴を総て捉えたことにはならず、最近ではより細かい単位での定量的な研究が進められている。

我々も、パラ言語・非言語情報と韻律的特徴との関連について、基本周波数パターンの面から生成過程モデルに基づいた研究を進めてきた。その中で、研究例が少ない、断定、疑問、反論といった意図と基本周波数パターンとの対応について調べ、特に文末でのアクセント成分の特徴が意図の識別に重要な役割を果たしていることを指摘している[3]。また、喜び、怒り、悲しみについて分析を行い、喜び、悲しみではアクセント成分の大きさが大きくなり、基本周波数の平均値が上昇することを示している。アクセント成分の大きさの増加は、基本周波数の偏差の増加に対応する。また、悲しみも含め、感情音声では平静の読み上げ音声に見られる文頭から文末に向かう基本周波数の減少傾向（declination）が見られなくなることを示している。

さらに、感情のレベルによって韻律的特徴がどのように変化するかを調べている[4]。模擬対話において文脈により怒り等の感情の度合いが平静を含め５段階に変化する状況を設定し、同一言語内容の発話の韻律的特徴の比較がなされている。図２は怒り、喜び、悲しみについて、生成過程モデルのアクセント成分の大きさ、フレーズ成分の大きさ、ベースライン周波数が、段階と共にどのように変化するかを示している。また、発話速度については、モーラ長の平静との差分を正規化したもので表現することを行なっている。分析結果が、各段階１例と少ないために、精密な議論ではないが、各特徴が段階の増加と共に一様に強調されているということにはなっていないことが推察される。武田ら[5]も、怒りについて段階による特徴の変化を調べ、段階の増加と共に各特徴が強調されるものの、激怒で発話速度の低下が見られることを指摘している。これらの結果から、我々は「人間はある感情を表現するのにいくつかの手段を持っているが、段階を表現する際、徐々にそれぞれの手段を顕著にしていくのではなく、利用する手段、あるいは割合を変更する。」と推測しているが、これは、感情表出の手段に大きな個人差があるということに対応していると言えよう。

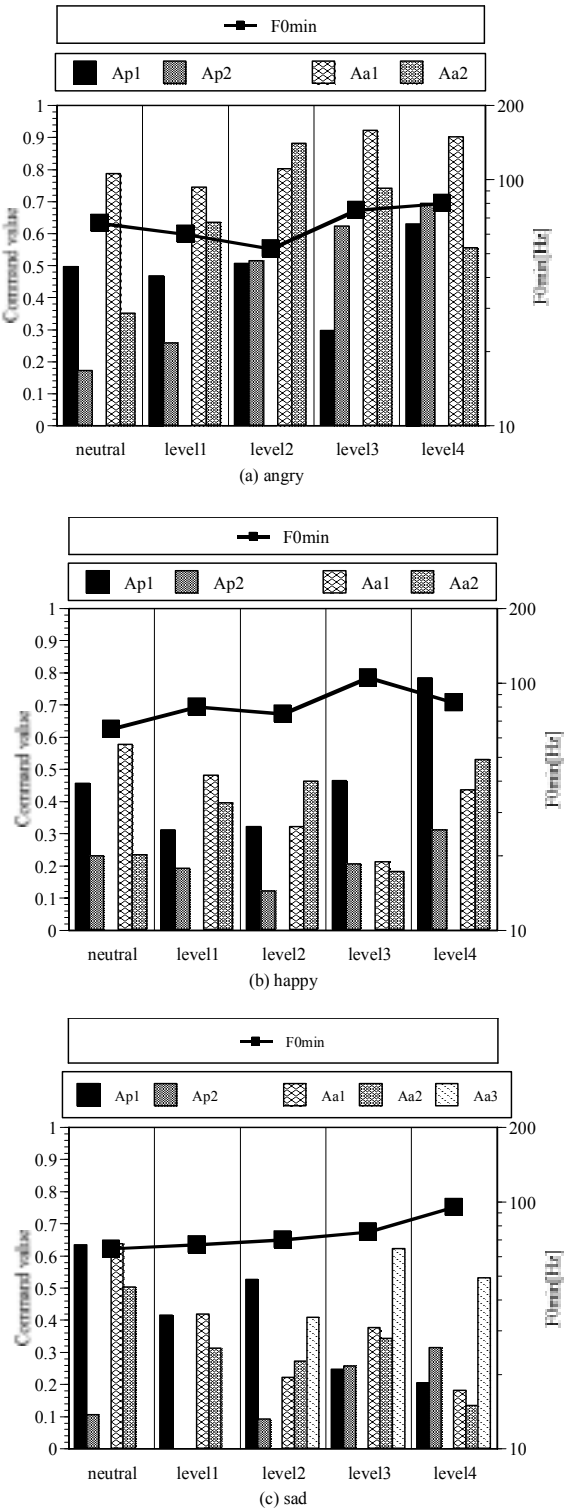

Figure 1.　F_0 model command values of ì ekimade mukaeni ikimasuî uttered by a male speaker in 4 levels of 3 types of emotions.

感情音声の特徴としては、スペクトルの特徴も重要である。スペクトルの傾斜、ピークの位置等についての議論は当初から行なわれていたが、最近、Kienastら[6]は母音のフォルマント、無声音のスペクトルのバランスについて詳しく調べている。怒りや喜びではtense、悲しみではlaxの特徴が現れるということに対

応した結果が得られている。最近では平館・赤木[7]によるHot AngerとCold Angerを対象とした分析もあり、フォルマント周波数が平静と比べ、前者では大きく、後者では小さくなるとの結果が得られている。スペクトルの特徴は、パワーと関連しても重要と考えられる。パワーについては、喜びや怒りでは大きく、悲しみでは小さくなるなどの特徴が見られ、感情の表現に重要な役割を果たしているが、これを実際に感情音声の合成、認識に利用するとなると問題がある。パワーは相対的なものであり、例えばマイクを近づけてゲインを上げればパワーは大きくなる。パワーが異なる発声では、音源波形形状の違いから当然、スペクトルのバランスも異なることになる。このように、パワーをそのまま取り扱うことは困難で、スペクトルの観点からの分析が必要となる。

3. 合成

基本感情を合成音声で表出することの研究は早くから行なわれているが、ルールベースで基本周波数、発話速度の全般的な傾向を記述し、制御するものが多い[8, 9]。問題は、基本周波数、発話速度といった制御の容易な特徴のみによって、どの程度、感情の合成が可能かということである。合成では、認識の場合と異なり、特徴の変動には対処する必要はないが、識別に対する寄与が大きくない特徴でも無視できないことが多い。個々の特徴を変更した合成音の聴取により、合成に際してどの特徴を制御する必要があるかが調べられている。我々も[4]、平静音声の音響的特徴を感情音声のそれで置き換えた音声を合成し、５段階評価で意図通りに感情が表現されているかの聴取実験を行なった。置き換えを、基本周波数のみ、音素持続時間のみ、パワーのみ、上記３者、ケプストラム係数（分節的特徴）のみの５通りについて行なった結果、表２のような結果が得られた。それによると、感情音声の合成に、分節的特徴が、特に喜びで重要である。

Table 2. Results of perceptual experiments (11 subjects).

Copied acoustic feature	Anger	Joy	Sadness
F_0 contour	1.3	1.6	2.2
Duration (speech rate)	1.1	1.2	1.3
Power	1.6	1.2	1.6
Three prosodic features	3.6	2.4	3.0
Cepstral coefficients	3.5	4.0	2.9

音響パラメータの詳細な制御が可能なフォルマント合成器を用いた合成が 1989 年頃から試みられている。Murrayら[10]は、DECTalkを基本としたHAMLETというシステムを開発し、ある程度の品質でそれとわかる 6 種の感情音声を合成している。最近では、Burkhardtら[11]が知覚的に詳細な実験を行なって感情音声合成のパラメータを求めている。

フォルマント合成は音響的特徴の柔軟な制御が可能な優れた方法であり、将来的にはルールベースの感情音声合成として、感情音声データが存在しない話者の感情音声の実現が期待される。しかしながら現時点では得られる合成音声の品質に問題がある。このため、接続型合成で感情音声の合成が試みられるようになり、実際、英語でのdiphone合成を始めとして多くの研究例がある[9]。最近では、感情音声データベースの整備により、コーパスベースの波形選択合成で感情音声の合成が行なわれるようになっている。飯田らはCampbellによるCHATRを感情音声データベースに適用することで、感情音声合成システムChatakoを構築している[12]。

波形選択合成では、TD-PSOLA によって素片の基本周波数等の韻律的特徴を修正し、感情音声に対応したものとする事が多い。この場合、問題となるのが、感情音声では韻律的特徴のダイナミックレンジが大きいために、修正量が大きくなり、音質劣化の要因となることである。Vineら[13]は、録音時にレファレンスとなるハム音を提示することにより、高低２種の基本周波数で素片を用意し、TD-PSOLA の修正量を小さくすることを行なったが、余り効果はないとのことであった。これは、2 種の素片のスペクトルも異なってくるためと考えられる。笠松ら[14]は、感情音声から１つの VCV 辺り４通りの素片を選択して用意することにより、感情が高精度に知覚可能な合成音声を実現している。

最近では、基本周波数パターンの生成に、統計的手法が導入されることが多くなっている。我々も、テキスト音声合成用に開発した生成過程モデルの枠組みに基づく基本周波数パターン生成手法を、感情音声の合成に適用していい結果を得ている[15]。この方法は、種々の言語情報を入力として、回帰木等の統計的手法により生成過程モデルの予測値を出力するものであり、生成過程モデルの制約があるために、大きな破綻がないといった特徴がある。

波形選択合成では、音質の良い合成音声が得られるが、相当量の感情音声コーパスが必要になる。感情をつけて文を読み上げることは、多くの人にとって必ずしも容易なことではなく、また、読み上げたとしてもそれは「朗読調の感情」という自然性に乏しいものになっている危険性がある。データなしは無理としてもなるべく小さな感情音声データベースから感情音声の合成を行なうことが求められる。このためには、多量の音声データが得やすい平静音声の特徴を感情音声に変更して合成する技術が必要となる。波形選択合成では、基本周波数等の変更は可能であるが、スペクトルの変更は困難である。前述したように感情音声の合成にはスペクトルの操作が必要である。これは、分析合成の枠組みで実現することができる。音声認識の分野

で成功した HMM を利用する合成方式が、徳田・小林らにより提案されており、適応技術を利用することで、少量のデータで音質等の変更が可能なことが示されている。これを用いて感情音声の合成が可能である[16]。

４．認識

4.1 パラ・非言語情報の識別・認識

ここ数年、感情を識別・認識する試みが数多く行なわれるようになっている。一般的には、基本周波数、持続時間、パワーといった韻律的特徴を特徴量として、判別関数、ニューラルネットワーク、ガウス混合分布を構築し、識別・認識する。Dellaert ら[17]はスプライン関数で補間した基本周波数パターンから得られる韻律に関する種々のパラメータ等を用いることにより、喜び、悲しみ、怒り、恐れについて人間と同程度の識別性能を達成している。音声認識に一般的な複数状態の HMM を用いることも行なわれ、刀根ら[18]は迷いと焦りという限定された対象ではあるが、良好な識別結果を得ている。ここで興味あるのは、話者によって識別率が大きく異なる点である。感情の認識は特定話者を対象とした場合が多く、不特定話者を対象とすると識別率が大きく低下する[19]。合成では問題となったスペクトルの特徴があまり用いられていないのはひとつにはその話者依存性が強いことに起因する。感情音声で重要な特徴は平静からの偏差であり、これを良好に表現する統計的な枠組みが必要であろう。

4.2 音声認識

朗読音声を学習用データとして構築された音素 HMM を用いて、朗読以外の音声を認識すると大きく認識率が低下する。感情音声を認識対象とした場合もそうであり、認識エンジンとして Julius を用いた場合、数十%の音素正解精度の低下が見られるとの報告がある[20]。何らかの適応が必要であるが、適応のためのデータが得にくいという問題がある。これに対して、我々は不特定話者モデルから 2 段の MLLR により、ごく少数のデータから特定話者の感情音声モデルを得る手法を開発し、認識率の向上を達成している[21]。

５．まとめ

感情音声の合成・認識にもコーパスベース手法が導入されるようになり、いかにして「作り物でない感情」の大規模コーパスを得るかが問題となりつつある。役者によるリストの読み上げは、多量の制御された音声が得られるという点からよく用いられているが、得られるのはあくまでも模擬した感情音声である。一般の会話を長時間録音することも行なわれているが、内容を制御できないため、思ったようなデータを得るのが困難である。

パラ言語・非言語情報は対話の円滑な進行に大きく寄与する。ここでは取り上げなかったが、例えば、発話権のやり取りに有効である。まだ話しつづけたい場合と、相手に応答してもらいたいときでは、明らかに韻律的特徴が異なってくる。今後、こういったことを取り扱いうる対話システムの開発が期待される。また、感性情報を対話システムで取り扱う場合、顔の表情等と組み合わせたマルチモーダル化が重要となる。顔の表情と組み合わせることで、より正確に感情を伝達することが可能となる。一方で、音声で表現された感情が表情に表れたものと異なる場合、非常に違和感があることが知られている。

感情音声で重要なのは、韻律的特徴であることは論を待たず、それに焦点を当てた研究が今後も必要である。この様な観点から、現在、筆者らが進めている文科省特定研究「韻律と音声処理」でも、パラ言語・非言語情報が重要テーマの１つとして取り上げられており、今後のその成果が期待される。

参考文献

[1] R. Cowie: Describing the emotional states expressed in speech, *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[2] R. Cowie et.al.: Emotion recognition in human-computer interaction, IEEE SP Magazine, pp.32-80, 2001-1.

[3] 広瀬啓吉: 音声コミュニケーションにおける感性情報, *感性の科学*, サイエンス社, pp.94-98, 1997.

[4] K. Hirose, N. Minematsu and H. Kawanami: Analytical and perceptual study on the role of acoustic features in realizing emotional speech, *Proc ICSLP*, Beijing, pp.369-372, 2000.

[5] 武田昌一他: 日本語「怒り」音声の韻律的特徴の解析, *日本音響学会誌*, 2002.

[6] M. Kienast, W. Sendlmeier: Acoustical analysis of spectral and temporal changes in emotional speech, *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[7] 平館郁雄, 赤木正人: 怒りの感情における音響特徴量の分析, *電子情報通信学会技術研究報告*; SP2002-141 pp.43-50, 2002.

[8] Y. Kitahara and Y. Tohkura: Prosodic control to express emotions for man-machine speech interaction, *IEICE Trans. on Fundamentals*, Vol.E75-A, No.2, pp.155-163, 1992.

[9] M. Schr ̂der: Emotional speech synthesis: A review, *Proc. EUROSPEECH*, Aalborg,, pp.561-564, 2001.

[10] I. Murray and J. Arnott: Implementation and testing of a system for producing emotion-by-rule in synthetic speech, *Speech Communication*, Vol.16, pp.369-390, 1995.

[11] F. Burkhardt and W. Sendlmeier: Verification of acoustic correlates of emotional speech using formant synthesis, *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[12] 飯田朱美他: 対話支援のための感情音声合成システムの思索と評価, *ヒューマンインタフェース学会誌*, Vol.2, No.2, pp.169-176, 2000.

[13] D.S.G Vine and R. Sahandi: Synthesising emotional speech by concatenative multiple pitch recorded speech units, *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[14] 笠松正紀他: 適応素片を用いた感情音声の合成, *電子情報通信学会技術研究報告*; SP2000-165, pp.41-46, 2000.

[15] 桂聡哉他: 生成過程モデルと統計的手法を用いた感情音声の基本周波数パターン生成, 音響学会全国大会講演集, 1-10-18, 2002-9.

[16] 田村正統他: HMM に基づく音声合成におけるピッチ・スペクトルの話者適応, 電子情報通信学会論文誌, Vol.J85-D-II, No.4, pp.545-553, 2002.

[17] F. Dellaert et. al.: Recognizing emotion in speech, Proc. ICSLP, Philadelphia, pp.1970-1973, 1996.

[18] 刀根優子他: 音声対話システムのための HMM に基づく感情判別, *電子情報通信学会技術研究報告*; SP2000-22, pp.47-53, 2000.

[19] L. Bosch: Emotions: What is possible in the ASR framework, *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[20] 門谷信愛希他: 音声に含まれる感情の判別に関する検討, *電子情報通信学会技術研究報告*; SP2000-82, pp.43-48, 2000.

[21] B. Li, et. al.: Robust speech recognition using inter-speaker and intra-speaker adaptation, *Proc. ICSLP*, Denver, 2002.