表現豊かな発話音声のテキスト評価と音声聴取結果に関する検討*

○金森康和，小川大介（愛知県立大），ニック・キャンベル（ＡＴＲ）

1　はじめに

音声は人間のコミュニケーションにとって不可欠の重要な方法の 1 つである。 最近、人々とコンピュータの間のインタフェースとして音声を利用する機会が増加した。さらに、それに加え、より人間に近い機能として、声にこめられた感情や感性をも音声認識や理解、合成に用いようとする研究が多くなされてきた。

過去の研究では、各種の情緒的な場面についてそれをテキストで記述した。被験者（発話者）がいくつかの希望する感情や発話スタイルを目指して、このテキストを音声で表現して、音声資料として使用された。しかし、それは通常のコミュニケーション[1-2]と異なる。 これらの理由で、より実際のコミュニケーション場面に近い音声資料の利用が好ましい。いくつかの報告書では、感情音声や表現豊かな発話音声 [3-5]に関する研究について報告された。N. Cambell and D. Erickson [4] は、クロス言語学において、この問題についての被験者の知覚に関して議論された。また、単語(「えっ」)の口語音声が様々な状況において日本語表現が使用された。その他に、「人々は何を聞くか。」[4] ということは音声研究者にとっての究極の課題であるかもしれない。

我々が前回の報告[5]では、データ数が少ないことが分かった。また、文意情報による影響があることも分かった。本報告では、我々の研究対象は主に全体の文である。人が音声信号から受ける発話の印象とテキスト文から受ける印象とが異なる。ここで、テキストのみを見た時の印象と、発話を聞いた時の印象について予備実験で作成された発話スタイルから選択することで印象の違いを調べる。この印象を調査すると共に、いくつかの音響特徴パラメーターが発話スタイルの判別分析に使った。

2　音声資料

本報告で用いる音声資料は論文 [3]で記述された自然な電話会話で、JST/CRESTデータベースの一部である。

日本語話者男性一人および女性一人によって発された 21 の対話を使用した。男性話者からの 12 対話、女性話者から 9 つの対話である。平静とは異なる様々な発話様式を考慮して、文の選択を行った。音声聴取の予備実験を通して、取りえる印象は苦笑い、本気笑い、愛想笑い、joking、驚き、疑問、納得、あきれ、困惑、考え中、興奮、好奇心、関心有、関心無、急いだ、穏やかな、ウワベと平静をあわせて 18 種類とした。以上のような方法で 147 の文が選択された。また、実験の重複性を見るために、以上から適当に 5 文が選ばれ、あわせて 152 文を実験に用いる。選択された各文の音声の長さは 1 秒から 20 秒に分布している。よって、文の長さは様々である。

3　特徴パラメーター

音声信号から抽出される基本周波数とエネルギーに注意を払うことが一般的である。本研究は前回[5]と同様、まず文中の基本周波数 F0 とエネルギーとしての RMS 値について、その平均値、最大値、最小値、レンジを特徴パラメーターとした。分析条件は表 1 と示す。

そのほかに、時間構造として、各文の発話時間長、有声音の発話速度（有声音の発話時間／有声音のモーラ数）、ポーズの長さ、モーラ数を用いた。また、基本周波数 F0 の傾き：/A, /B, /C, /D, /E, |/E|も用いた[5]。

表 1　F0 と RMS の分析条件

分析条件	値
サンプリング周波数	16 kHz
ビット･レート	16ビット
フレーム長さ	16ミリ秒
フレーム・シフト	10ミリ秒

* Study about text evaluation and speech listening results for the expressive expression speech, by KANAMORI Yasukazu, OGAWA Daisuke (Aichi Prefectural University) and Nick Campbell (ATR).

4　音声聴取実験とテキスト文実験

2節で述べた音声資料（152 文）をランダムに提示し、男性 12 名と女性３名を含む被験者が、18 種類の発話様式から知覚されたものを答える。各文の聴取に対して、この 18 スタイル中から多重選択を許可し、また、度合い２と度合い１で答える。任意の発話様式に対して、強く知覚されたときは度合い２を、単に知覚されたときは度合い１を回答してもらった。ただし、本報告ではある様式が知覚されたかどうかが重要であると考えた。すなわち、程度から発生する差は分析しない。なお、これらの被験者は大学生である。実験は防音室でヘッドホンによって実施した。

5 つの重複文について、15 の被験者が、各々がほとんど同じ様式にマークしている。15 人の結果が安定性上信頼できると考えられる。

上述の音声を書き起こしたテキスト文を 15 人中の 10 人に同様な実験を行った。ただし、今回はランダムな順序でプリントしたテキスト文のみを見て回答する。

5　結果と考察

各文について、多種な様式が選択されることがあるので、ここでは、被験者全員の評価で得点の多い様式をこの文の様式としている。Fig.1 にはテキストのみからの印象と音声聴取で得た印象が同じ様式となった文の割合(一致率)を示す。音声信号とテキスト文から受ける印象が被験者によって大きく異なる。その一致する度合いは 18.4%から 60.5%まで分布する。

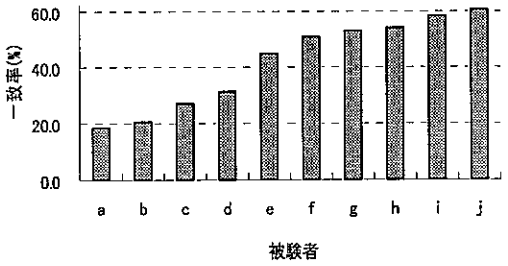

Fig.1 被験者によるテキストと音声からの印象が一致する度合い

また、特徴パラメーターを用いて、３文以上を持つ音声の発話様式を判別すると、表２のような結果が得た(笑い文を除く)。これはテキスト文と音声からの印象が一致するとき、テキストの文意からも様式が推定でき、文意情報からの影響が大きいといえる。そこで、文意影響を考慮して印象が不一致する文(59 文)について判別実験をすると、その様式判別率は 66%(97 文中)から 80%(59 文中)に上昇した。

表２　判別実験の結果

	文意影響を考慮		音声＋文意		一致数
	数	的中率	数	的中率	
joking	3	100. 0%	3	100. 0%	0
驚き	9	66. 7%	12	41. 7%	3
疑問	9	77. 8%	17	58.8%	8
納得	17	82. 5%	29	65.5%	12
考え中	4	50.0%	15	73.3%	11
興奮	4	100.0%	7	85.7%	3
関心(有	3	100.0%	3	100. 0%	0
関心(無	3	100.0%	4	75. 0%	1
ウワベ	4	50.0%	4	50. 0%	0
平静	3	100.0%	3	66. 7%	0
計	59	79.7%	97	66.0%	

6　まとめ

本報告では、音声信号からの印象とテキスト文から受ける印象がどのように被験者に影響するかを調べた。結果として被験者よって大きく異なる。２つの印象が一致する度合いは 20%から 60%まで分布する。また、テキスト文と音声からの印象が不一致の文について、発話の時間構造、F0 の傾き、F0 と RMS 値の平均値や最大値などの特徴パラメーターを用いて判別すると、その様式判別率は 66%(97 文中)から 80%(59 文中)に上昇した。

参考文献

[1] E. Krammer, “Elimination of verbal cues in judgments of emotion from voice,” Journal of Abnormal and social Psychology, Vol. 68, pp.390-396 (1964)

[2] K. Maekawa, “Phonetic and phonological characteristics of paralinguistic information in spoken Japanese,” Proceedings of ICSLP, pp.635-638 (1998)

[3] N. Cambell, “JST/CREST ESP Project: Expressive Speech Processing”, Proceeding of The 1st JST/CREST International Workshop on Expressive Speech Processing,(IWESP) Kobe, Japan, pp61-70 (2003)

[4] N. Cambell and D. Erickson, “What do people hear? A study of the perception on non-verbal affective information in conversational speech,” Journal of the Phonetic Society of Japan, Vol.8 No.1, pp.9-28(2004)

[5] 市川、金森, “自然発話音声における感情の分析”、音講論集, 1-7-16, pp.243-244(2004-3)