論　文

自由発話音声における音響的な特徴の検討

村上　仁一[†]　　嵯峨山茂樹[††]

A Discussion of Acoustic Problems in Spontaneous Speech Recognition

Jin'ichi MURAKAMI[†] and S. SAGAYAMA[††]

あらまし
本論文では、自由発話の特徴について朗読発話と比較して調べた結果を述べる。自由発話の音声認識は今後の大きな課題であるが、従来の朗読発話の認識と比較すると様々な困難が予想される。そこで視察によるラベリングをして自由発話と朗読発話の差を調べた。調査に用いたデータは、4 名のナレータによる対話音声である。この結果を以下に示す。
（1） 文章全体の 40%から 65%の文に、間投詞が出現する。
（2） 間投詞の出現頻度の 50%から 75%は、4 種類の間投詞で占める。
（3） 文章全体の 2%から 4%の文に、言い直しが出現する。
（4） 発話速度は、最も差がある話者でも 6%しか増加しない。
（5） 融合ラベルの付与率は、約 20%も増加する話者がいる。
（6） 連続音素認識の実験の結果、認識精度 (Phone Accuracy) は 6%から 10%しか減少しない。
キーワード
自由発話, 発話速度, 間投詞, 言い直し, 音素認識率

1. まえがき

近年、連続音声認識の研究が盛んに行なわれ、多くの研究機関で文音声音声システムが構築されている [1], [2], [3], [4]。これらのシステムの多くは、朗読発話のような丁寧に発声された音声を入力対象にしている。しかし、人間同士のコミュニケーションでは、「あのー」「えーと」などの間投詞や、言い淀みや言い誤りおよび言い直しなどが頻繁に見受けられる。このような音声の認識が今後の重要な研究課題になると思われる。

この研究の第一歩として、本論文では視察によるラベリングをして自由発話の音声データを調べた。自由発話の定義は研究者によって異なるが、ここでは話者がテキストを見ないで対話した音声を自由発話と見なした。そして自由発話と朗読発話の差を見るために、間投詞と言い直しの出現頻度、発話速度、融合ラベルの付与率、HMM による認識精度などを調査した。ただし、調査した話者は 4 名のみであるため、調査結果の値は一般性に欠ける可能性がある。

なお、自由発話の視察によるラベリングには多くの人手が必要であるため、このような報告は少ない。文献 [5] において、小林らは日本音響学会のデータベース [6] を利用して、自由発話の文の中に出現するポーズの長さを報告している。一方、音声を文字化するコストはラベリングのコストよりも少なくてすむため、間投詞や言い直しなどの言語現象を調べた論文は比較的多い。日本語では、間投詞や言い直しの出現頻度を調べた報告 [7], [8], [9] や、助詞落ち・倒置の分析を行なった報告 [10] などがある。英語では自由発話のデータベースとして Air Travel Information Service (ATIS) がよく知られている。このデータベースを利用して自由発話の特徴を報告し [11], [12]、従来の音声認識で使用されたアルゴリズムを用いて、認識率を報告した論文が多く見られる [13], [14]。

[†] ATR 音声翻訳通信研究所, 京都府
ATR Interpreting Telephony Reseach Labs. , Kyoto-fu, 619-02 Japan
[††] NTT ヒューマンインターフェース研究所, 神奈川県
NTT Human Interface Labs., Yokosuka-shi, 238-03 Japan

2. 自由発話の特徴

2.1 調査に用いたデータベース

現在ＡＴＲでは、各種言語現象を調査するために対話文を中心とする言語データベースの作成を進めている [15]。本来、対話音声の収録は話者に録音していることを気づかれずに録音することが好ましいが、通信の守秘義務などの問題の他に、話題が次々に移行するため会話の語彙が膨大な数になるという問題も生じる。このため、事前に話題のトピックやバックグラウンドを決め、会議の流れの不自然さを損なわないように打合せを行った後に収録をしている [16]。現在、発話内容で 5 種類、収録環境で 2 種類、話者で 2 種類、発話様式で 2 種類の variety を含むデータベースを収集中である [15]。

本論文では、このＡＴＲの言語データベースの中から、申し込み者と事務局員が国際会議の問い合わせを想定して集録した音声データの中の 4 名の話者を、自由発話の音声データとして利用した。この音声データは、ナレータ (アナウンサや声優など音声を職業としている人) が申し込み者の役になって発話しているため、舌打ちの音などはほとんどない。また、事務局員側はこの分野の専門家が演じているが、この音声データは調査に用いていない。両話者は完全に分離されて録音されているため音声区間の重畳はない。 収録は遮音室で行なわれたためドアの開閉音などの日常雑音はない。したがって、この音声データは自由発話としてはかなり clean な音声であると言ってよい。このデータの収録条件を表 1 に示す。ただし、各々の話者の発話内容は異なっている。

表 1 調査に用いた自由発話の音声データの収録条件
Table 1 Recording conditions for spontaneous speech data

話者	ナレータ 2 名 (通称 MTK,FKN)
収録環境	遮音室
発話内容	国際会議の申し込みに関する参加者と事務局の対話
発話様式	自由発話 「トピック」(質問項目と、その背景に関する情報) や「バックグラウンド」(会話の前提になる背景) を詳細に設定して対話したもの。
入力系	マイクロフォン、ＤＡＴ録音
データ量	13 対話 116 文 3943 音素 (MTK) 18 対話 333 文 11520 音素 (MMY) 13 対話 180 文 6918 音素 (FKN) 14 対話 195 文 7588 音素 (FAK)

また、自由発話の音声データを文字に書き起こした後に、間投詞や言い直しを削除して作成したテキストを、自由発話と同一の話者が発声した音声を、朗読発話の音声データとして使用した。したがって、同一話者における自由発話と朗読発話の発話内容は、間投詞および言い直しを除いてほぼ同一である。

なお、自由発話では文の定義が曖昧になる。本論文では、発話権が相手から渡されて相手に返すまでの話者の発話を「文」と定義した。したがって、通常、複数の文と考えられる文が 1 文になる。発話の例文を表 2 に示す。この例文において“『　』”で括られた範囲は 1 文を示している。また“『”の前の番号は、文番号である。

表 2 文　　例
Table 2 Examples for Spontaneous Data

1 『[あっ、あの] わたくし、 武蔵野電機システム開発研究所の小金沢と申します。』
2 『[あのー] 参加の申込みをしていたんですけれども、 [ちょっと] 出られなくなりましたので、 [あの] キャンセルしたいんですが。』
3 『はい。[あのー] キャンセルは書面にてっていうふうに 書いてあるんですけども、 どういうふうにしたらよろしいんでしょうか。』
4 『[ああー]。』
5 『[あっ] そうですか。』
6 『[あっ] そうですか。』
7 『それでですね、 [えーと] ８０パーセント返していただけるのは 9月３０日までというふうに書いてあるんですけども、 [あのー] ３０日までにそれが届けばですか。
そ　れとも、こちらが出した [その] 消印が ３０日でも (かまい) かまわないんでしょうか。』
8 『[ああ] そうですか。』
9 『[あっ] そうですか。』
１０『わかりました。じゃ、登録ナンバーと名前、住所、所属なんかを (かき) 書いて、書いたものをお送りすればよろしいですね。』
１１『はい、どうも、ありがとうございました。』

2.2 間投詞および言い直しの出現頻度

自由発話には、「あのー」や「えーと」などの間投詞 (冗長語) や、言い直しがある。これらの出現頻度を各話者ごとに調べた。

なお、本論文では間投詞を『自立語。活用しない。主語・述語によらない。言い淀む場合などに、文の中に挿入されて用いられる。間投詞を取り除いても文の文法性および意味には影響しない [17]。』と定義した。また言い直しを『前にいった事の誤りを訂正してもう一度言う。』もしくは『他の適当なやさしい言葉で言う。』と定義した。また言い淀みを『言おうとしてためらったり、話しの途中でちょっと言葉につまったりする。』[18] と定義した。なお表 2 の例文において、

“[ ]”で括られた個所は間投詞、“( )”で括られた個所は言い直しを意味している。

各話者ごとの間投詞の出現頻度を図 1 に、言い直しの出現頻度を図 2 に示す。

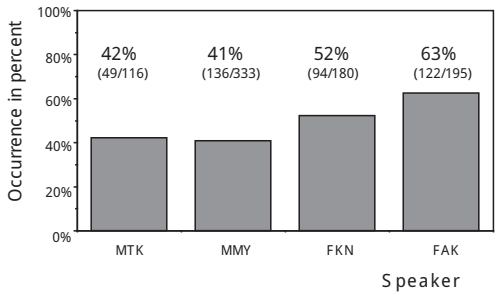

図 1　間投詞の出現頻度
Fig.1　Occurrence of filled pauses

図 1　間投詞の出現頻度
Fig. 1　Occurrence of filles pauses

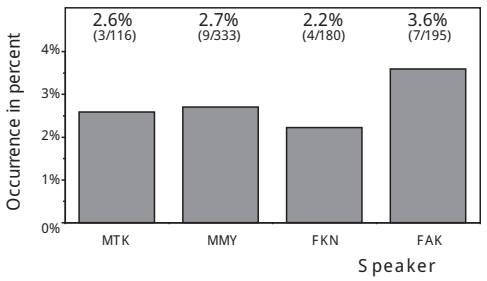

図 2　言い直しの出現頻度
Fig.2　Occurrence of self-corrections

図 2　言い直しの出現頻度
Fig. 2　Occurrence of self-corrections

これらから次のようなことが判る。

（１） 間投詞は、話者によって相違が見られるが、文章全体の 40%から 60%の文に出現する。

（２） 言い直しは、話者によって相違が見られるが、文章全体の 2%から 4%の文に出現する。

2.3　間投詞の種類と出現頻度

間投詞には多くの種類があるが、出現頻度の高い間投詞は限られていることが既に報告されている [7], [5], [24]。ここでは各話者ごとに、出現頻度の高い上位４つの間投詞の種類と、その出現頻度を調べた。この結果を図 3 に示す。

これらから次のようなことが判る。

（１） 間投詞全体の 50%から 75%は、使用頻度の高い４種類の間投詞で占める。

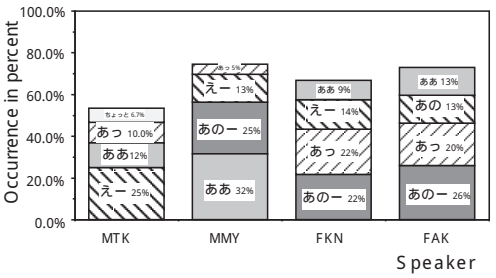

図 3　間投詞の種類と出現頻度
Fig.3　Kind of filled pauses

図 3　間投詞の種類別の出現頻度
Fig. 3　Kind of filled pauses

（２） 使用頻度の高い間投詞の種類は、話者によって相違がある。

2.4　ラベリング作業からみた自由発話

音声データのラベリングの作業において見受けられた自由発話の音素の定性的な特徴を表 3 に示す。なお、ラベリングの基準は文献 [19] に従った。これらから自由発話では音素境界がかなり曖昧になっていることや、従来の朗読発話には見られない音素が現れていることがわかる。

表 3　自由発話の音素の定性的特徴
Table 3　Qualitative Characteristic for Spontaneous Speech

1	文の語尾の音素の発音が弱くなることがある。 （例: 「なんですか」の「か」がほとんど聞こえない。） （ただし、発声が弱くでも音素境界が明瞭な場合がある。）
2	母音/a,i,u,e,o/全てが無声化することもある。 （朗読発話では/a,e,o/は、あまり無声化しない。） （例：「それで」の/o/と/e/が無声化する場合がある。）
3	判断できない不明瞭な音素がある。 （例:「んー」（考え込むとき発声している音）は /N/あるいは/uN/の 両者に解釈できる。）
4	子音/r/をともなう音節の発音が全体的に弱い。 （例:「そうす る と」の「る」がほとんど聞こえない。）
5	母音（特に、文末の母音『ａ』）の第１フォルマントが あらわれないことがある。

2.5　融合ラベルの付与率から見た自由発話

ＡＴＲでは、発話テキストを参照しながら人手で音素境界を決定するラベリング作業において、音素境界が不明瞭な音素区間に対して付与するラベルのことを融合ラベルと呼んでいる。この融合ラベルの付与率を 4 名の話者の自由発話と朗読発話において調査した。この結果を図 4 に示す。

この図から読みとれることを以下に示す。

（１） 自由発話の融合ラベルの付与率は朗読発話より高い。

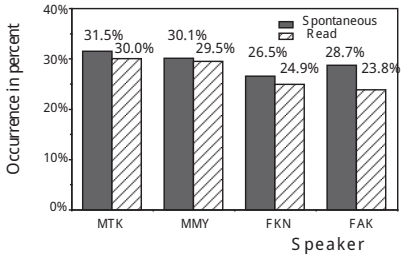

図 4　融合ラベルの付与率
Fig.4　Frequency of compound phone labels
(Labels that could not further be divided into isolated phone labels)

図 4　発話様式の違いによる融合ラベルの付与率の変化
Fig. 4　Frequency of compound phone labels

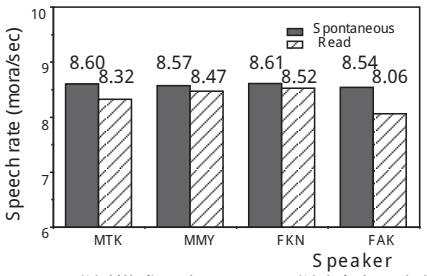

図 5　発話様式の違いによる発話速度の変化
Fig.5　Speech rate variance depending on speaking style

図 5　発話様式の違いによる発話速度の変化
Fig. 5　Speech rate variance depending ons speaking style

（２）自由発話、朗読発話共に、融合ラベルの付与率に話者の相違が見られる。

（３）自由発話では、全音素の 25%から 32%が融合ラベルになる。

（４）朗読発話では、全音素の 24%から 30%が融合ラベルになる。

（５）自由発話と朗読発話を比較すると、融合ラベルの付与率の増加の割合に話者の相違が見られる。話者 MMY では 2%(29.5% → 30.1%) しか増加しないのに対し、話者 FAK では 21%(23.8% → 28.7%) 増加する。

なお、 /i,y/, /u,g/, /N,g/, /N,j/, /f,u/, /i,m/, /k,u/, /u,h/, /u,w/, /u,y/, /u,n/, /N,b/, /N,d/, /N,n/, /a,a/, /e,e/, /e,i/, /i,i/, /k,i/, /o,o/, /o,u/, /s,u/, /sh,i/, /u,u/ などの音素環境は朗読発話、自由発話ともに、融合ラベルになりがちであった。また、文末の 2 音素を調査したところ、 母音では/e/と/u/、子音では/g/と/n/が融合ラベルになりやすかった。

2.6　発話速度からみた自由発話

ここでは、自由発話と朗読発話の発話速度の差をモーラ速度で調査した。ただし、息つぎなどの長いポーズ区間および間投詞および言い直しの音声区間は除去した。この結果を図 5 に示す。

また、自由発話および朗読発話における各音素の平均音素継続時間を表 4 に示す。ただし融合ラベルが付与された音素は評価対象から削除した。

これらの結果から読み取れることを以下に示す。

（１）自由発話の発話速度は朗読発話より早い。

（２）自由発話の発話速度は 8.5(mora/sec) から 8.6(mora/sec) である。

（３）朗読発話の発話速度は自由発話より話者の相違が大きく 8.1(mora/sec) から 8.5(mora/sec) である。

（４）話者 FAK を除くと、自由発話における母音/a/の平均音素継続時間は朗読発話より短い。しかし母音/u/,/e/,/o/は朗読発話より長い。

（５）話者 MTK を除くと、自由発話の音素継続時間の分散は朗読発話より大きい音素が多い。

表 4　各音素ごとの平均音素継続時間
Table 4　Duration Times for Each Phone

自由発話

話者	MTK		MMY		FKN		FAK	
音素	平均 (ms)	標準偏差	平均 (ms)	標準偏差	平均 (ms)	標準偏差	平均 (ms)	標準偏差
a	92.6	49.1	87.6	41.0	86.1	38.2	93.4	63.4
i	72.0	36.1	65.1	39.1	64.2	36.8	70.5	45.5
u	85.0	62.0	77.0	53.8	69.0	43.4	77.7	45.5
e	93.5	64.3	92.7	72.2	80.6	44.3	85.0	53.7
o	91.8	61.5	97.4	71.6	92.5	59.6	101.3	74.5
p	61.4	7.1	17.4	6.9	20.7	7.8	12.8	4.5
t	41.4	21.0	20.4	7.1	17.2	5.7	14.8	5.8
k	49.3	19.9	36.3	19.3	30.2	10.5	26.6	12.1
b	48.3	13.4	11.1	5.0	11.8	4.4	11.7	2.4
d	43.5	16.9	12.0	4.7	11.6	4.0	10.8	3.3
g	45.9	25.3	13.8	4.7	13.9	5.5	14.3	5.2
m	53.1	17.3	47.1	19.6	52.6	15.6	51.8	17.2
n	48.2	18.3	41.6	20.0	47.9	19.1	53.4	24.6
N	75.7	24.0	66.6	37.5	55.2	28.2	70.7	32.4

朗読発話

話者	MTK		MMY		FKN		FAK	
音素	平均 (ms)	標準偏差	平均 (ms)	標準偏差	平均 (ms)	標準偏差	平均 (ms)	標準偏差
a	100.2	57.5	88.7	36.3	88.0	33.3	91.5	43.1
i	81.3	46.9	70.8	36.3	63.6	30.9	70.7	35.1
u	76.1	48.6	71.2	42.0	66.8	39.4	73.3	38.5
e	92.2	67.3	85.6	58.6	73.6	28.6	82.3	41.8
o	86.9	50.2	91.7	53.5	86.5	38.5	94.5	46.7
p	61.3	14.1	17.6	7.4	16.8	4.5	17.9	8.9
t	40.1	22.9	18.6	6.7	15.3	4.6	13.8	4.3
k	42.5	18.7	35.0	18.1	28.1	10.5	28.1	12.8
b	52.9	38.9	9.7	2.2	11.6	5.1	10.0	0.0
d	43.1	18.7	11.7	4.1	11.3	3.1	10.0	2.5
g	45.6	26.7	19.7	9.8	13.6	5.7	13.5	4.0
m	53.4	18.1	48.1	19.0	47.2	14.2	50.1	14.6
n	47.3	15.5	44.5	21.7	44.1	17.6	48.4	16.1
N	103.6	57.4	70.2	33.4	63.3	37.7	65.3	24.1

なお、話者 MMY における音素/a/と/k/の継続時間毎の分布図を図 6 に示す。この図では、横軸は音素の継続時間で、縦軸は出現回数である。この図から、音素/a/では自由発話の音素継続時間の分散は朗読発話より大きいことがわかる。しかし音素/k/では継続時間の分散の差が小さいことがわかる。

2.7 認識精度 (Phone Accuracy) から見た自由発話

ここでは自由発話と朗読発話の差を、連続音素認識実験をして正解率 (Phone Correct) および認識精度 (Phone Accuracy) [21], [20] で評価した。特定話者の同一発話様式の認識実験を行なうために、同一話者の同一発話様式の音声データの、文番号の奇数番目を学習データに偶数番目を評価データにした。学習プログラムには主に HTK Software Tools [20] を使用した。特徴パラメータには LPC ケプストラムを使用し、HMM には混合連続分布型を用いた。表 5 に実験条件を示す。

表 5 音素認識の実験条件
Table 5 Experimental Conditions for Phone Recognition Rates

認識対象	26 音素
サンプリング周波数	12kHz
話者	男性のナレータ
学習データ	同一発話様式
音響パラメータ	log power + 16 次 LPCcepstrum + Δlog power + 16 次 Δcepstrum
フレーム窓長	20ms
フレーム周期	5ms
LPC 分析	16 次
打ち切り次数	16 次
音素モデル	4-state 3-loop 3 mixture Gaussian continuous HMM (diagoal)

認識実験は以下のようにしておこなった。

（１） 学習データにおいて、融合ラベルが付与されなかった音素のみを切り出して Baum-Welch 学習をする。学習回数は１０回。

（２） 学習データを文単位で連結学習する。学習データは間投詞や言い直しを含む。学習回数は３回。

（３） 学習データと同一話者・同一発話様式の評価データを文単位で連続音素認識する。なお評価データは間投詞や言い直しを含む。

（４） 評価データの音素ラベルを正解として、音素正解率 (Phone Correct) と音素認識精度 (Phone Accuracy) を計算する。

2.8 音素認識実験から見た自由発話

図 7 に、認識実験の結果得られた音素正解率 (Phone Correct) と音素認識精度 (Phone Accuracy) を示す。また母音の音素認識誤り傾向を表 6 に示す。

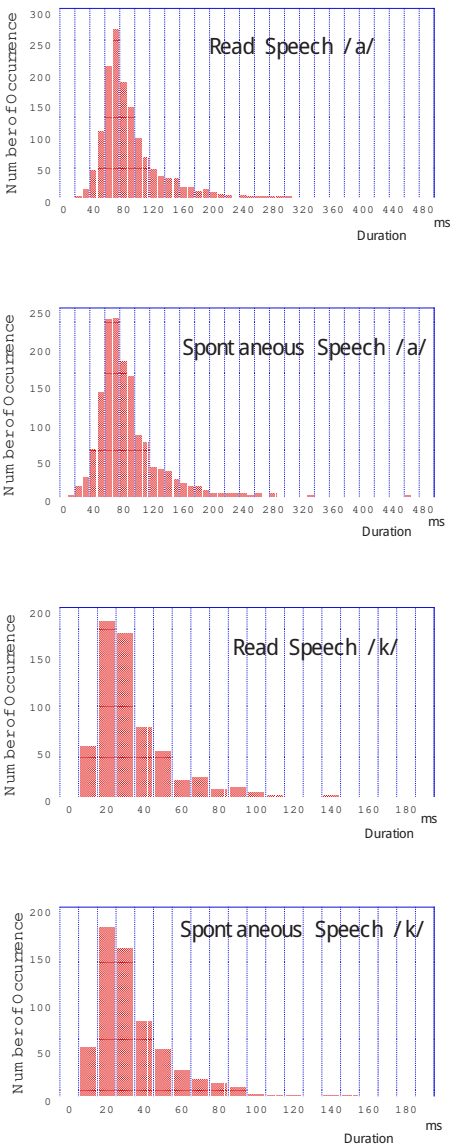

図 6 音素/a/および/k/の継続時間の分布 (話者MMY)
Fig.6 Distribution Map of Duration Length for /a/ and /k/ (Speaker MMY)

図 6 継続時間の分布
Fig. 6 Distribution map of duration length of /a/ and /k/

これから次のような結果が示される。

（１） 自由発話は朗読発話と比較して、音素正解率も音素認識率も低下する。

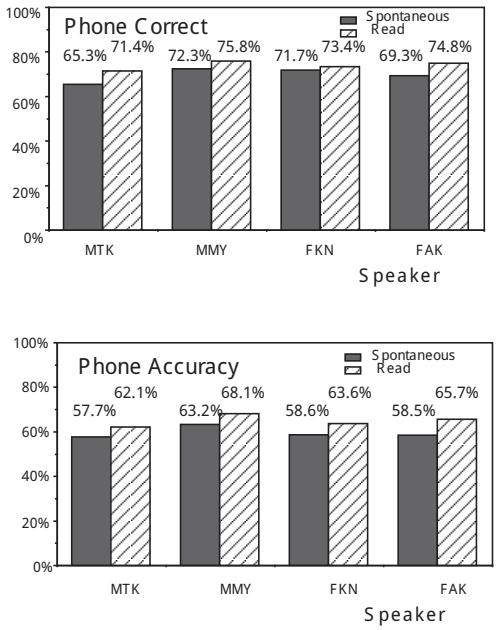

図7　音素認識実験結果
Fig.7　Result of experiments

図 7　音素認識率 (%)
Fig. 7　Result of experiments

（２）自由発話の正解率 (Phone Correct) は、65%から 72%が得られた。

（３）自由発話の認識精度 (Phone Accuracy) は、58%から 63%が得られた。

（４）自由発話は朗読発話と比較すると認識精度は 7%から 10%程度低下する。

（５）各音素の認識率をみると、母音の/u/の認識精度が他の音素と比較して低い。

3. 考　　察

3.1　間投詞の出現頻度と種類に関して

今回の調査では、話者によって相違があるが、間投詞が出現する文は文章全体の 40%から 65%を占めることが示された。しかし、電話のような音声のみによる対話では、間投詞は相手の注意を促す役割を持つ場合がある [8]。したがって話し相手と対面して話す自由発話では、この出現頻度より低くなる可能性がある。

なお、自由発話における間投詞（冗長語、不要語）の出現頻度は多くの研究期間で報告されている。文献 [22] や文献 [23] では日本音響学会連続音声データベースの書き起こしテキストを調査して報告している。また、文献 [8] では、開始符合としての間投詞の種類と出現頻度を報告している。また、文献 [24] では NHK ラジオ第一放送の電話相談番組を書き起こして報告している。これらの論文と比較すると、間投詞の出現頻度はほぼ同じ割合と言える。また、間投詞の種類も、これらの報告と比較すると比率に違いがあるが、代表的な間投詞に関してはほぼ同じ割合といえる。

表 6　音素認識誤り傾向
Table 6　Tendancy for Phone Recognition Errors

話者 MTK (認識音素数/調査音素数)

入力＼出力	a	i	u	e	o
a	83.1% (167/201)	0.0% (0/201)	1.5% (3/201)	3% (6/201)	7.5% (15/201)
i	0.7% (1/128)	85.1% (109/128)	3.9% (5/128)	3.9% (5/128)	0.7% (1/128)
u	7.3% (6/ 82)	4.8% (4/ 82)	48.7% (40/ 82)	2.4% (2/ 82)	9.7% (8/ 82)
e	3.0% (4/131)	13.7% (18/131)	1.5% (2/131)	76.3% (100/131)	2.2% (3/131)
o	3.5% (5/140)	0.7% (1/140)	5.7% (8/140)	2.8% (4/140)	80.0% (112/140)

話者 MMY (認識音素数/調査音素数)

入力＼出力	a	i	u	e	o
a	93.2% (633/679)	0.1% (1/679)	0.7% (5/679)	3.0% (21/679)	1.1% (8/679)
i	0.0% (0/426)	81.4% (347/426)	3.2% (14/426)	4.9% (21/426)	0.0% (0/426)
u	1.2% (4/320)	4.0% (13/320)	45.6% (146/320)	3.4% (11/320)	7.1% (23/320)
e	1.4% (6/405)	3.4% (14/405)	2.2% (9/405)	83.4% (338/405)	0.7% (3/405)
o	1.5% (8/522)	0.0% (0/522)	1.7% (9/522)	3.4% (18/522)	88.5% (462/522)

話者 FKN (認識音素数/調査音素数)

入力＼出力	a	i	u	e	o
a	83.7% (381/455)	0.4% (2/455)	1.9% (9/455)	4.6% (21/455)	1.5% (7/455)
i	0.0% (0/289)	76.4% (221/289)	2.0% (6/289)	3.8% (11/289)	0.3% (1/289)
u	1.4% (3/205)	0.9% (2/205)	52.6% (108/205)	9.7% (20/205)	4.3% (9/205)
e	0.4% (1/227)	4.8% (11/227)	3.0% (7/227)	84.1% (191/227)	0.0% (0/227)
o	1.2% (4/318)	0.0% (0/318)	4.4% (14/318)	0.3% (1/318)	88.6% (282/318)

話者 FAK (認識音素数/調査音素数)

入力＼出力	a	i	u	e	o
a	80.6% (393/487)	0.0% (0/487)	4.1% (20/487)	4.7% (23/487)	2.0% (10/487)
i	0.0% (0/265)	73.9% (196/265)	1.1% (3/265)	7.9% (21/265)	0.3% (1/265)
u	6.0% (12/199)	3.5% (7/199)	43.2% (86/199)	6.0% (12/199)	4.0% (8/199)
e	0.8% (2/244)	9.0% (22/244)	2.8% (7/244)	78.6% (192/244)	0.4% (1/244)
o	2.6% (10/381)	0.0% (0/381)	3.9% (15/381)	1.5% (6/381)	83.7% (319/381)

3.2　自由発話における言い直しに関して

今回の調査では、話者によって相違が見られるが、言い直しを含む文は文章全体の 2% から 4%を占めることが示された。しかし、今回の調査した話者はナ

レータ (アナウンサや声優など音声を職業としている人) であるため、一般の人の言い直しの出現頻度は、これよりも高いと思われる [7]。

なお、文献 [25] では言い直した単語に着目して、言い直しを分析している。これを見ると言い直しの 59%は、言い誤った単語を直ちに言い直している。また文献 [26] においてもほぼ同様な結果が見られる。今回の自由発話データの言い誤りを分析すると、単語にならない音節となっているものが 39%、直後に言い直しているのが 52%であり、傾向はほぼ同じであった。

3.3 自由発話と朗読発話の音響的な差

本論文では、自由発話の音響的な特徴を調査するために、主に融合ラベルの付与率、発話速度、HMM における音素認識誤り率で朗読発話と比較した。その結果、自由発話は朗読発話と比較すると、発話速度は最も差がある話者でも 6%しか増加しないが、融合ラベルの出現頻度は約 20%も増加する話者がいることが示された。しかし、自由発話と朗読発話の認識精度 (Phone accuracy) の差は 7%から 10%程度であることが示された。したがって、少なくとも同一話者 (特定話者)、同一発話様式で HMM を学習をする限り、音響モデルに関しては自由発話と朗読発話に大きな差はないように思われる。

ただし、本論文で調査した話者は音声による対話に慣れた人である。したがって、一般の話者が雑音下で制約の少ない状態で話した音声では、この論文で調査した結果と異なる可能性がある。

3.4 自由発話の可能性について

自由発話において特徴的な言語現象に、間投詞や言い直し・言い誤り・言い淀みなどがある。そして、今回の調査の結果、間投詞は発話全体の 40%から 65%の文に、言い直しは約 2%から 4%の文に出現することが示された。自由発話の認識には、これらの言語現象の処理方法が大きな問題になると考えられる。

現在自由発話の認識アルゴリズムとしては、これらの現象に対応するため、1) キーワードスポッティングを利用する方法 [27]、2) 音素モデルにガーベージモデルなどを使用して認識する方法 [28] [29]、3) 言語モデルの一部に音素系列として認識する方法 [30], [22] もしくはこれらの組合せの手法 [13] などが試みられている。しかし、これらのアルゴリズムには挿入誤りが増加することや、広いビーム幅が要求されるなどの問題点が残っている。

4. ま と め

ここでは自由発話の認識にむけて、4 名のナレータにおける自由発話の音響的な特徴を調べた。この結果、話者によって相違が見られるが、間投詞が出現する文は文章全体の 40%から 65%を占めることや、間投詞全体の出現頻度の 50%から 75% は 4 種類の間投詞で占めることがわかった。また、言い直しは文章全体の 2%から 4% に出現することがわかった。

そして、視察による音素ラベリングの結果、自由発話は朗読発話と比較すると、発話速度は、最も差がある話者でも 6%しか増加しないが、融合ラベルの付与率は約 20%も増加する話者がいることがわかった。しかし連続音素認識の実験の結果、認識精度 (Phone Accuracy) は朗読発話と比較すると 6%から 10%しか減少しないことがわかった。したがって、これらの点を考慮すると、間投詞や言い直しなどの言語現象を除けば、少なくとも音素モデルに関しては、同一話者 (特定話者)、同一発話様式において音素の HMM の学習をしたならば、自由発話と朗読発話に大きな差はないように思われる。

ただし、ここで扱った自由発話は、言葉の対応に慣れた人たちが限定した条件の下で発話したデータである。したがって、一般の話者が、雑音下で制約の少ない状態で話した音声では、この論文で調査した結果と若干異なる可能性がある。

謝辞

研究で使用した自由発話の言語のデータベースの作成は、匂坂氏や現ＮＨＫの江原氏の指示のもとになされたものです。また、ATR 音声翻訳通信研究所の森元室長や飯田室長の他、各研究員に多くの協力をいただきました。これらの方々に厚くお礼を申しあげます。

文　献

[1] Kai-Fu Lee, “Large-Vovabulary Speaker Independent Continuous Speech Recognitton: The SPHINX System,” Computer Science Department Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 CMU-CS-88-148 (April 18, 1988).

[2] 永井 明人, 他, “HMM-LR 連続音声認識装置の開発と性能評価”, 日本音響学会平成 3 年度秋季研究発表会, 1-5-23, pp.45-46, (Oct. 1991).

[3] Xuedong Huang, et., “An Overview of the SPHINX-2 Speech Recognition System”, ARPA Human Langauge Technology Workshop, (Mar. 1993)

[4] Rechard schwarts, et., “Comparative experiments on large vocabulary speech recongition”, ARPA Human Langauge Technology Workshop, (Mar. 1993)

[5] 小林 聡, 山本 幹雄, 中川 聖一, "間投詞・言い直し等の出現に関する音響的特徴", 情報処理学会, 音声言語処理研究グループ資料 93-SLP-1-2, pp.7-10 (1993).
[6] 田中 和世, 板橋 秀一, 他, "音声の知的処理に関する調査報告書", システム技術開発調査研究 3-R-2, 財団法人 機械システム振興協会, (平成４年３月)
[7] 村上 仁一, 嵯峨山 茂樹, "自由発話音声認識における音響的、言語的な問題点の考察", 信学技報, SP91-100, pp.71-78 (Dec. 1991).
[8] 飯田 仁, 野垣内 出, 相沢 輝昭, "通訳を介した電話対話の特徴分析", 電子通信学会技術速報, NLC86-11 (1986).
[9] 有田 英一, 小暮 潔, 野垣内 出, 飯田 仁, "メディアに依存する会話の様式," NL61-5 (1987).
[10] 山本 幹雄, 小林 聡, 中川 聖一, "音声対話文における助詞落ち・倒置の分析と解析手法", 情報処理学会論文誌 Vol.11, No11, pp.1322-1330 (1992).
[11] Victor Zue, James Glass, David Goodine, et al. "The MIT ATIS System: Preliminary Development, Spontaneous Speech Data Collection,And Performance Evaluation", Eurospeech 91, pp.537-540, Genova, Italy, (Sep. 1991)
[12] Victor Zue, Nancy Daly, et al, "The Collection and Preliminay Analysis of a Spontaneous Speech Database", DARPA Workshop 1989, pp.126-134 (1989).
[13] Wayne Ward and Sunil Issar, "Recent Improvedments in the CMU Spoken Language inderstanding system", ARPA HUMAN LANGUAGE TECHNOLOGY WORKSHOP, pp.208-211, (Mar. 1994).
[14] Victor Zue, Et al " Pegasus: A Spoken Language Interface for On-Line Air Travel Planning", ARPA HUMAN LANGUAGE TECHNOLOGY WORKSHOP, pp.196-201, (Mar. 1994).
[15] 江原 暉将, 小倉 健太郎, 森元 逞, "電話対話データベースの構築", 第 40 回情報処理全国大会, pp.486-491 (1990).
[16] 篠崎 直子, 小倉 健太郎, 森元 逞, "言語データベース作成のためのシュミレーション会話", 第 37 回情報処理全国大会, pp.1000-1001 (1988).
[17] 吉本啓, " 日本語品詞の分類", ATR Techinical Report TR-I-0008 (Nov,1987).
[18] 金田一京助, 他, "新明解 国語辞典 第３版", (1987-10).
[19] 武田 一哉, 匂坂 芳典, 片桐 滋, 桑原 尚夫,"音韻ラベルの持つ日本語音声データベースの構築", SP87-19 (1987).
[20] Cambridge University Engineering Department Speech Group and Entopic Research Laboratories Inc., "HTK:Hidden Markov Model Toolkit V1.5" (23 Sepember 1993).
[21] 中川聖一, "音声入力を想定したあいまいな発話文の理解システムに関する研究", 文部省科学研究費補助金, 一般研究 (B) 研究成果報告書, pp.137-144, (平成６年３月) .
[22] 小林 聡, 甲斐、山本 幹雄, 中川 聖一, "間投詞の出現位置の特徴分析と音声認識システムの評価", 情報処理学会, 音声言語処理研究グループ資料, 92-SLP-3-4 (1992).
[23] 高木 一幸, 保浦 直子, 板橋 秀一, "対話における話題展開と発話単位の性質", 情報処理学会, 音声言語処理研究グループ資料, 93-SLP-1-3 pp.11-18 (1993).
[24] 佐川雄二、大西昇、杉江昇、 "対話文における誤りの自動修復", 情報処理学会 自然言語処理研究会資料, 93-10, pp.71-78 (Jan. 1993)
[25] E.Shriberg, J.Bear and J.Dowding,"Automatic detection and corrction of repairs in Human-Computer Dialog", Proc. Speech and Natual Language Workshop, pp.419-424 (1992).
[26] C.Nakatani and J.Hirschberg, "A Speech-First Model for Repair Detection and correction", In Proceedings 31st Annual meeting of theAssociation for computational linguistics, pp.46-53 (1993).
[27] 坪井 宏之, 橋本 秀樹, 竹林 洋一, "連続音声理解のためのキーワードラティスの解析," 日本音響学会平成３年度秋季研究発表会, 1-5-11, pp.21-22 (Oct. 1991).
[28] 井ノ上, 武田 一哉, 山本, "Garbage HMM を用いた自由発話文中の不要語処理手法", 電子情報通信学会論文誌,A Vol.J77-A, No.2, pp.215-222 (1994.2).
[29] 今村 明弘、北井幹雄, "事後確率を用いたフレーム同期型ワードスポッティング," 日本音響学会平成５年度秋季研究発表会, 1-4-2, pp.3-4 (Mar. 1993).
[30] 南 泰浩, 山田智一, 鹿野清宏, "番号案内を対象とした大語彙連続音声認識アルゴリズム", 電子情報通信学会技術速報, SP92-108 (Dec. 1992).

Abstract

In this paper, we present a preliminary study of spontaneous speech recognition, describing the acoustic characteristics of spontaneous speech.

Recognition of spontaneous speech is a hard problem which is known to be considerably more difficult than recognition of read speech. To get a better understanding of spontaneous speech and how it differs from read speech, a preliminary study has been carried out to compare spontaneous and read speech utterances for 4 broadcast announcers.

In spontaneous speech, the speech rate increased by a maximum of 6%. When labeling spontaneous speech, the labeling uncertainty increased by a maximum of 20% as compared to read speech. In phoneme recognition experiments we found a decrease of between 6% and 10% in the phoneme recognition accuracy for spontaneous speech.