テレビ番組音声の自発的発話音声コーパスとしての活用可能性について

高橋　　徹†，赤塚　俊洋††

Applicability of Speech Sounds in TV Programs for Spontaneous Speech Corpus

TAKAHASHI Toru†，AKATSUKA Toshihiro††

Abstract

Since a large-scale speech corpus is required to train the spontaneous speech recognition engine, it is required to support the collection of speech data. We show that it is possible to collect spontaneous speech sounds from sounds in television programs. The collected sound shows applicability as speech corpus for training the spontaneous speech recognition engine. As a result of investigation of the sounds in TV program, we conclude that spontaneous speech sounds can be collected efficiently from the sounds in variety programs.

キーワード：Speech Corpus，Speech Recognition，Sounds in television program

1　はじめに

音声認識技術は，年々より自然な発話音声を扱えるようになってきている。孤立単語認識[1)]の研究から始められ，定型文[2,3)]，非定型文[4)]と文法的な制約の緩和という流れを辿ってきた。同時に特定話者の音声認識から不特定話者[5,6,7)]へと話者性に関する制約も緩和されてきた。単語発話[1)]，連結した単語列[2)]の発話，連続発話[8)]，読上げ発話[9)]，講演発話[10)]，演技発話，

†大阪産業大学　デザイン工学部　情報システム学科准教授
††大阪産業大学大学院　工学研究科　情報システム工学専攻
草稿提出日　12月16日
最終原稿提出日　1月12日

表１：コーパスサイズの変遷

年代	コーパスサイズ	例
1990年	～10時間	研究者独自に個々に作成
	10時間～	JNAS[15]（読上げ音声）
2000年	100時間	ATR-BLA[16]（読上げ音声）
	100時間～	CSJ[17]（話し言葉音声）
2010年	1000時間～	衆議院審議[18]
	5000時間～	Google Voice Search

自発的発話[11]と話し方の多様性にも対応されてきた。

認識対象音声の自由度が増加し，識別すべきパターンの多様化を引き起している。その結果として，認識機は，膨大なパターンを扱う必要性が生じている。現在，一般的な認識機は，隠れマルコフモデルに基づいた手法である[12, 13, 14]。音声の特徴量を瞬時特徴とその時間変化として表現し，そのパターンを識別する手法である。このパターンは音響特徴量と呼ばれ，具体的には，多次元ベクトル系列で表される。音響特徴量は，機械学習によって統計的に処理され，大量の音響特徴量から音声の特徴を分布として獲得する。この分布を音響モデルと呼び，入力音声は，音響モデルに基づき確率的に認識される。

機械学習に基づき統計的に認識エンジンを構成するには，多くの学習データが必要となる。認識対象となる音声の制約が緩いほど，発話の多様性が広く，それらを網羅する学習データが必要となる。自然な会話音声を認識するためには，膨大な学習データを要する。

学習データは，音声とその発話内容のスクリプトの組が大量に登録されたデータベースで音声コーパスと呼ばれる。表１に示すように，音声コーパスサイズは，多様な音声を扱う研究に発展するに連れ増加し続けている。不特定話者による自由文法の自発的発話を認識するためには数千時間ものコーパスが使われるようになった。例えばGoogle Voice Searchでは数千時間から数万時間のコーパスによって認識エンジンが学習されている。今後，非母国語話者による発話を認識するなどといった，より多様な発話を認識するためには，更に膨大なコーパスが必要になる。この様に必要なコーパスを用意することが困難になっているという課題がある。

自発的な話し言葉を扱う認識エンジンを構成するために必要となる膨大な音声コーパスを容易に収集する方法が必要である。Google Voice Searchでは，日々の検索クエリーを学習データに追加し，認識エンジンを再学習することで認識精度を改善できる。しかし，この方法は検索事業を広く展開している特定の企業のみが実現できる手法である。そこで我々は，大量の音声データを収集する手段としてテレビ番組の音声に着目した。

テレビ放送は，10チャネル程度の番組がほぼ24時間365日放送されている。ケーブルテレビを含め地域の違いも合せると10チャネル以上の番組が随時放送されている。これらの音声を集めると，単純計算で１年で87,600（=10*24*365）時間収集できる。実際には，87,600時間の録

音音声を全て音声コーパスに使用できるとは限らないが，仮に10%に限られたとしても，8,760時間の音声を収集できる。

コーパスとして利用できる音声は，発話者のみの音声が単独で流れている区間である。無音部分，楽曲が流れる部分，音声発話が楽曲や効果音と重なっている部分，複数の話者が同時に話している部分は音声コーパスに適さない。テレビ番組音声のうちコーパスとして利用できる音声がどの程度収集可能か調査する必要がある。

本研究では，音声コーパスに適した発話が，どのような番組に多く含むかを検討する。単独発話を多数含みかつ様々な話者が様々な発話スタイルで発声する番組が適している。 BGMや効果音が少ない番組ほど，音が重ならず単独発話を収集しやすいことが想定できる。その様な番組の特徴を調査する。また，テレビ番組中の音声が自発的発話と見做せる音響特徴の多様性を有するかを調査する。

２　テレビ番組中の音

テレビ番組中の音を分類し，番組中に現れる音源のうち，音声が単独で現れる割合を明かにすることが目的である。またテレビ番組中の音を用いることから，テレビ番組内で使用頻度の高い音源，つまり音楽と効果音も分類する。本実験では（１）音声，（２）音楽，（３）効果音他の３種類の音源を扱う。

２.１　調査対象の選択

調査対象とするテレビ番組の選択は，出演者数，BGMの有無，効果音の有無といった多様性を持つよう配慮した。最終的に，ニュース番組，アニメ，ドラマ，バラエティ番組を調査対象とした。ニュース番組は,ニュースキャスターによるニュース原稿を読上げるシーンが多く，比較的読上げ音声に近い音声が出現すると想定し選択した。ドラマ，アニメにはBGMが多く使用されていると考えた。俳優や声優が台詞を自発的に発話しているかのように演技して発話することから，発話スタイルを演技スタイルであると考えて選択した。バラエティ番組は，脚

表２：出演者，音楽，効果音の数

番組の種類	出演者数	音楽数	効果音数	発話スタイル
ニュース	4	1	14	読上げ
アニメ	6	3	42	演技
ドラマ	13	2	23	演技
バラエティ１	10	1	2	自発
バラエティ２	6	1	4	自発
バラエティ３	8	1	2	自発

本があっても台本はないと考えられる。仮に台本があったとしても一言一句台本の台詞を再現することが目的の番組ではないため，出演者の発話を自発的な発話として扱う。出演者数に幅があるためいくつかの番組を取り上げる。番組の選定結果を表２に示す。

２.２　音源の列挙

ニュース，アニメ，ドラマ，バラエティ番組中の音を音源ごとにその出現区間を調査する。録画した番組を人手により繰り返し視聴し調査した。番組中の音声発話（セリフ，ナレーション），音楽（BGM, 歌），その他（効果音, 笑いごえ, 周囲の物音など）を人が聞き分け分類した。更に，音源の出現区間を，音の開始点，終了点としてラベル付けした。

２.３　音の単独出現率と重複出現率

音の単独出現率 R_Sと重複出現率 R_Mは以下の式で求められる。

$$R_S = \frac{s}{T} \times 100\ [\%] \qquad (1)$$

$$R_M = \frac{m}{T} \times 100\ [\%] \qquad (2)$$

ここで，sは音が単独に現れる時間の合計，mは音が他の音源と同時に現れる時間の合計である。Tは，コマーシャル区間と無音区間を除いた番組長である。

２.４　調査結果

単独出現率と重複出現率を番組別に求めた結果を表３に示す。次に番組別に単独出現音の内訳を表４に示す。これらの結果から単位時間当たりに番組音から得られる単独出現音声の割合（利得）

$$\text{利得} = \frac{\text{単独出現の音声長}}{\text{番組長}} \qquad (3)$$

を求めた。ニュース番組が，音声の単独出現音率（利得）が最大であった。続いてバラエティ番組 １，３，２の順である。

２.５　考察

ニュースは，読上げ音声を認識するための音声コーパス素材収集に適している。ニュース原稿を読上げるスタイルであるため，単独発話が多く含まれている。利得も70%以上であり，高

表３：音の出現時間

番組	単独出現時間	重複出現時間
ニュース	1,298（s）	378（s）
アニメ	498（s）	888（s）
ドラマ	274（s）	599（s）
バラエティ１	627（s）	671（s）
バラエティ２	446（s）	1,349（s）
バラエティ３	639（s）	315（s）

表４：単独出現音の内訳

番組	音声	音楽	効果音他	利得
ニュース	1,222（s）	13（s）	63（s）	72.9（%）
アニメ	126（s）	230（s）	142（s）	9.1（%）
ドラマ	91（s）	127（s）	56（s）	10.4（%）
バラエティ1	526（s）	18（s）	83（s）	40.5（%）
バラエティ2	360（s）	27（s）	59（s）	20.1（%）
バラエティ3	450（s）	16（s）	173（s）	47.2（%）

い効率で音声素材を収集できる。しかし，ニュース発話は，読上げスタイルであり，自発的な発話として収集するには不適切である。

アニメは，演技発話を認識するための音声コーパス素材に適している。しかし，BGMや効果音が多用され，音声の単独発話区間が少なく，利得は９%程度である。ニュースに比べて利得が低く素材収集の効率が悪い。オープニングやエンディング音楽が必ず含まれることが単独発話時間の短さになっている。更に，効果音の占める時間も長いことも利得の低さになっている。

ドラマは，演技発話といってもアニメに比べて自然な発話に近い音声である。しかし，アニメと同様に利得が10%程度と低い。BGMをはじめとした音楽が多用されていることが影響している。

バラエティは，利得が20～40%と中程度である。ニュースと比較すると利得は低いが，自発的な発話の収集目的としては，十分な利得であると考えられる。当初の目的である自発的な発話を認識するための認識エンジンを学習するために収集する音声コーパス素材として適当である。バラエティ1，３は，ひな壇に出演者やゲストを並べ，司会者が進行する形式の番組である。特にこれらの利得は40%程度で効率が良い。

以上より，自発的な発話音を20～40%程度の効率で収集するには，バラエティ番組が適していると結論付けられる。特にひな壇形式の様な出演者が多く，司会が全体を進行する形式の番組を選択すると効率が良い傾向も確認できた。

3　バラエティ番組内の音声特徴

前章では，テレビ番組音声から単独発話音声を効率よく収集するためにはバラエティ番組音声が適することを確認した。本章では音声認識エンジンの構成について概説し，バラエティ番組内の音声が，どの程度自発的な発話として扱えられるかを音声認識エンジンを使って検証する。

3.1　音声認識エンジン

音声を認識するために必要な音声認識エンジンの構成について述べ，音響モデルと言語モデルが重要な役割を果すことを説明する。

はじめに発話された音声が，フレーム分析によって短時間スペクトルに変換される。20〜40msの短時間領域を短時間フーリエ変換し短時間スペクトルを得る。発話の始めから終わりまで，フーリエ変換する時間領域を５〜10ms進めながら変換を繰り返すと短時間スペクトル列が得られる。この短時間スペクトル列が発話の音響特徴である。一般には短時間スペクトル列からMFCC（Mel-Frequency Cepstral Coefficients）と呼ばれる低次元の特徴量に変換され，音響特徴量として使用されることが多い。

音声認識エンジンの役割は，音響特徴から発話の最小単位である音素の列を推定することと，得られた音素列から，言語や文法知識に基づき単語列を推定し，一つの文章として解釈することである。音声認識エンジンは，音素列を推定するために音響モデルを用い，音素列から文章を推定するために言語モデルを用いる。

最近の音声認識エンジンの音響モデルには，その認識性能の高さからHMM[12]やDNN-HMM[13, 14]が用いられている。これらは，時間周波数に広がる音響特徴の時間構造をHMMで表し，周波数構造を確率分布としてガウス混合分布やDeep Neural Networkで表している。音声は，特定の単語を速く発声することもゆっくり発声することも可能である。そのため同じ単語の音響特徴が，異なる時間構造を持ち得る。HMMはその時間伸縮をモデル化している。

例えば，未知の３音素からなる単語が/a/, /r/, /u/という音素列に認識されるまでを考える。全ての３音素並びの音素系列に対応するHMMと未知の音素列の音響特徴を比較する。尤も高い確率で未知の音響特徴を出力するHMMを選ぶ。選ばれたHMMが何の音素列を表すHMMであったかを確認する。/a/, /r/, /u/という３つの音素列から構成されたHMMであれば，未知の音響特徴が，/a/, /r/, /u/と確認される。最終的に入力音声が「ある」と発音されたものと解釈される。

問題は，時間周波数構造が，読上げ音声と自発的な発話で異なる点である。従って，認識エンジンで自発的な発話を認識するために，読上げ音声の音響モデルを使用すると音響モデルのミスマッチによって，認識精度が低下する。自発的な発話でよく見られる「発音のなまけ」や

「言い澱み」によって時間周波数構造の違いが現れミスマッチが発生する。

言語モデルも同様の問題がある。言語モデルは，連続するN個の語がその語順で出現する頻度を大量のテキストデータベースから算出し，特定の並びの語が出現する確率を表している。言語として出現する確率の高い語順になるよう優先的に認識するために用いられる。従って，自発的な発話を認識するために，認識エンジンで読上げ音声の言語モデルを使用すると言語モデルのミスマッチによって，認識精度が低下する。通常言語モデルは，大量のテキストデータベース，つまり書き言葉から語順の出現頻度を算出する。自発的な発話は，書き言葉と異なる語順を用いることが多くミスマッチが発生する。

3.2　自発的な発話の文法的な比較

ここでは読上げ音声と自発的な発話音声のミスマッチを使って自発的な発話の文法的な多様性を確認する。

自発的な発話の音声を人手で書き起こし，その発話内容を計算機による自動形態素解析する。人手で形態素解析した結果と比較する。文法に則った自発的な発話では，これら２つの解析結果が等しい。

自発的な発話に，何らかの言い澱みや文法から逸脱が見られると，解析結果は異なる。この時，自発的な発話に言い澱みや文法から逸脱があったと考えることができる。

この解析結果の差異を確認し，バラエティ番組のうち発話内容が文法に従っている割合を確認する。

バラエティ番組中に出現した単独音声の内，表５の22文を自動形態素解析機chasen[19] で解析したところ22文中21文が正しく解析可能であった。ただし，形態素解析が方言に非対応なため，方言部分を除いて解析結果の正しさを判定した。バラエティ番組内の発話は，文法に基づ

表５：単独発話の例

1	幾つになったの本当に	12	なるかわからないというやつやろ
2	言いたくない	13	なったんです
3	本気で知りたいです	14	あのね四つ選ばれた
4	知りたいというか	15	おもてなしやと思うよ
5	えーだいたい二十二三の頃	16	ほんとですか
6	えーよくお会いしてた頃がそうやろ	17	持って来たというか
7	経っちゃったんです	18	たまたまあれだけが
8	結婚してないよねまだ	19	日本語だったんですよ
9	こんなはずじゃなかった	20	みんなのプレゼンの中で
10	変わってないです	21	きっといいおもてなしをしますよ
11	色々活動して	22	さあ今日は久々か会うのが

かない発話が存在することを確認できる。正しく解析できなかった文は22番の文である。倒置と間投詞の挿入によって文法から逸脱している。この結果から，読上げ発話用の言語モデルでは，全ての自発的な発話へ対応可能ではないことを確認できた。

３.３　自発的な発話の音響的比較

本節は，自発的な発話における音響的特徴を考察する。文法的に読上げ発話と同等な特徴を持ち，音響的特徴だけが自発的な発話の特徴を持った自発的な発話を使う。この様な発話は，自発的な発話音声のうち，その書き起こしテキストが正しく形態素解析可能な発話の集合である。我々は，表５中の１～21文を用いて実験する。

収集した21文を，読上げ発話用の音響モデルの音声認識エンジンで認識処理すると，音素列を得る。この音素列を，人手で聞き取った音素列と比較する。もし，同じ系列が得られれば，その自発的な発話は，読上げ発話的に発声されたと解釈できる。異なる系列が得られれば，その自発的な発話は，読上げ発話から逸脱した自発的な発話特有な発声と解釈できる。

21文中21文が，読上げ音声用の認識エンジンでは認識できなかった。一方,これら21文を著者の１人が読上げた音声を収録し，読上げ音声用の認識エンジンで認識すると，正しい音素系列が得られた。本研究の目的は,自発的な発話特有な発声を大量に収集することである。従って，認識誤りのある文が自発的な発話の音響モデルの学習に適した発話である。これら21文の発話は,読上げ発話の特徴の分布と異なる音響的特徴を有していると結論付けられる。これは，真に自発的な発話に表れる特徴である。これらの発話を大量に収集し，機械学習によって自発的な発話の音響モデルを構築できる可能性が示唆された。

４　おわりに

本研究では，テレビ番組音声を自発的な発話のための音声認識エンジンを学習する用途に使用する可能性について検討した。またその収集効率の高い番組の特徴について調査した。単独発話の出現割合の多い番組は，出演者の多いバラエティ番組であった。音が存在する区間中20～40%程度の割合で含まれていた。効率的に自発的な発話収集に，バラエティ番組が適していることが確認できた。

バラエティ番組中の単独発話の中から,文法から逸脱していない発話を収集した。これらは，読上げ音声用の音響モデルで認識困難であることから，収集したバラエティ番組中の発話は，音響的に自発的な発話であると結論付けられる。

従って，本研究の目的通り，テレビ番組中の音声から音響的な自発的な発話音声を収集可能であることを示すことができた。また，収集音は，自発的な発話の音声コーパスに替わり，音声認識エンジンの学習データに活用できる可能性を示すことができた。

今後，テレビ番組音声中の単独発話区間を自動で抽出する信号処理アルゴリズムを開発し，自動的に自発的な発話音声を収集する方法を確立したいと考えている。

謝辞

本研究は，科学研究補助事業，基盤研究（C）15K00254，15K00255，大阪産業大学産業研究所分野別若手研究平成25～27年度の支援のもとに行われた。

参考文献

1）L. Rabiner, J.Wilpon, “Isolated word recognition using a two-pass pattern recognition approach,” *Acoustics, Speech, and Signal Processing, IEEE*, pp.724–727, 1981.
2）L. Rabiner, C. Schmidt, “Application of dynamic time warping to connected digit recognition,” *IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.28, Issue.4*, pp.377–388, 1980.
3）J. L. Gauvain, J. Mariani, “A method for connected word recognition and word spotting on a microprocessor,” *Acoustics, Speech, and Signal Processing, IEEE, vol.7*, pp.891–894, 1982.
4）J. Markel, S. Davis, “Text-independent speaker recognition from a large linguistically unconstrained time-spaced database,” *IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.27, Isuue.1*, pp.74–82, 1979.
5）Marvin R. Sambur, “Speaker Recognition Using Orthogonal Linear Prediction,” *IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.24, Issue.4*, pp.283–289, 1976.
6）B. S. Atal, “Automatic recognition of speakers from their voices,” *IEEE Journals and Magazines*, pp.460–475, 1976.
7）L. Rabiner, S. E. Levinson, A. E. Rosenberg, J. G. Wilpon, “Speaker-Independent Recognition of Isolated Words Using Clustering Techniques,” *IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No.4*, pp.336–349, 1979.
8）F. Jelinek, “Continuous speech recognition by statistical methods,” *IEEE Journals and Magazines*, pp.532–556, 1976.
9）L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis, R. Mercer, “Speech recognition of a natural text read as isolated words,” *Acoustics, Speech, and Signal Processing, Vol.6*, pp.1168–1171, 1981.
10）西村他，“講義コーパスを用いた自由発話の大語彙連続音声認識，”電子情報通信学会論文誌，D-II, Vol.J83-D-II No.11, pp.2473–2480, 2000.
11）Y. Minami, K. Shikano, S. Takahashi, T. Yamada, “Search algorithm that merges candidates in meaning level for very large vocabulary spontaneous speech recognition,” *Acoustics, Speech, and Signal Processing, Vol.2*, pp. 141–144, 1994.
12）B. H. Juang, “On the hidden Markov model and dynamic time warping for speech recognition —A unified view,” *AT&T Bell Laboratories Technical Journal, Vol.63, Issue.7*, pp.1213–1243, 1984.
13）D. Yu, L. Deng, G. Dahl, “Roles of pretraining and finetuning in context-depenent DBN-HMMs for real-world speech recognition,” *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

14）Jinyu Li, Dong Yu, Jui-Ting Huang, Yifan Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM,” *IEEE spoken language technology workshop*, pp.131–136, 2012.
15）伊藤，他，ASJ（E），“JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research”, Vol.20, No.3, pp.199–206, 1999.
16）ATR音声データベース：*http://www.atr-p.com/products/sdb.html*
17）CSJ音声コーパス：*http://pj.ninjal.ac.jp/corpus center/csj/*
18）河原達也，議会の会議録作成のための音声認識―衆議院のシステムの概要―，情報処理学会，研究報告，音声言語情報処理，Vol.93, No.5, pp.1–6, 2012.
19）形態素解析ソフトウェア，chasen: *https://ja.osdn.net/projects/chasen-legacy/*