表現豊かな自然発話コーパスのアクセスについて

モクタリ 明子† 田畑 安希子† ニック・キャンベル‡
†神戸大学大学院総合人間科学研究科 ‡ダブリン大学トリニティカレッジ

本稿では，2000年から2005年までの5年間に渡って収録した大規模自然発話音声コーパス（全収録時間1,500時間以上）の一部を紹介する．コーパスの話者は学生・教員・主婦・子供など様々であり，全ての発話が課題なしの自然発話である．現在このコーパスを公開するための作業を進めているWEBページでは，書き起こしテキストを目で追いながら該当する発話音声を聞くことや，特定の語彙を含む発話を検索することが可能である．また，同コーパスのうちすでに研究利用されている他のデータについても，その内容および研究成果を紹介する．
キーワード：自発的発話，インターネットアクセス，発話様式，外国語教育

Access to an Expressive Speech Corpus

Akiko Mokhtari† Akiko Tabata† Nick Campbell‡
† Graduate School of Intercultural Studies Kobe University
‡School of Linguistic, Speech and Communication Sciences Trinity College Dublin

This paper introduces a part of a large natural spontaneous conversation corpus (total recording time is more than 1,500 hours) which was collected over the period of five years from 2000 to 2005. Speakers are ordinary people including students, professors, housewives and children, and no task was given to the speakers. Part of the data is available on a website which is periodically updated, where you can listen to utterances with an aligned display of the speech annotation, and where you can also search a particular expression from the data base. This paper also describes other data of the corpus and some of the research results obtained from these data.
Keywords: Spontaneous speech, internet access, speech style, foreign language education

1．はじめに

言語研究や音声認識技術の向上など様々な目的のため幅広い研究分野において音声コーパスの需要は年々高まっている．いずれの場合においてもコーパスをデザインする際に，いかに本来我々が自然な状態で発している発話音声に近いものを，音質を落とすことなく収集できるかということは大きな課題であろう．例えばこれまで感情や態度に伴う音声の変動について，主に音声工学や実験音声学において多くの研究がなされてきた．これらの研究の中には，悲しい物語や楽しい物語などを話者に読ませることによって，できる限り自然な感情・態度が込められた音声を引き出そうとしているものや（Iida et al.: 2003），飛行機事故の実況中継を伝えるラジオのアナウンサーから得られた恐怖などの感情が込められている音声を分析対象とした研究もなされている（Stevens & Williams: 1972）．またErickson et al.（2004）は実験室でモノログ的に収録されたものではあるものの，スクリプトなどは用意せず，自発的発話を通して偶発的に収録することができた悲しみの発話音声を刺激音として用いている．しかし依然としてアナウンサーによる朗読音声を基本とする「実験室的環境」で収録されたデータが大部分の研究において用いられているのが現状である．

実験の主旨や研究の背景となる考え方によって，適切なデータの収録方法は異なってくるため，一概にどの方法が最も適切であるかを指摘することはできない．しかし現実の日常コミュニケーションと実験室的環境で収録された音声データとの間に大きな隔たりがあることが多くの研究者によって意識されているのも事実である（e.g. Cowie: 2000, Maekawa et al.: 2000）．

我々は日々の生活の中で，発話の背後にある意図を表したり，テキストだけでは伝えられない情報を付け加えたりするために音色や発話様式を変化させている．このような表現豊かな発話音声は，コントロールされた実験室的環境で収録されたデータでは到底カバーしきれないものである．

そこで日常の発話音声がもつ表現豊かな音色を反映する対話音声を収集するために自然環境で1,000時間の音声データを収録することを目指したのがESPプロジェクトである．ESPプロジェクトは，ニック・キャンベルをプロジェクトリーダーとし，2000年から20005年までの5年間に渡り，科学技術振興機構（JST）の支援を受け，大規模自然発話コーパス（以下，ESP

コーパス）を構築した．このプロジェクトは，コントロールされたデータの収集や信号処理を行うのではなく，データ数を増やすことで日常会話の中に現れる豊かな音色をカバーすることができるという考え方のもとに進められた．収録に参加した話者は全て一般の人であった．収録された音声データは 2 人以上の話者によるインタラクティブな対話であり，収録場所はデータによって異なるものの話者の自宅や大学の研究室など話者がリラックスして話せる場所が多い．後述するように，より高音質のデータを収録するために録音ブースで行われた対話も一部含まれているが，いずれの場合も話者に課題は一切与えられず全てのデータが完全に自発的な発話である．

ESP プロジェクトの研究実施体制は，研究項目の異なる 7 つのサブグループから構成された．データ収集は主に神戸大学国際文化学部の教員・同大学院総合人間科学研究科の院生から成る意味構造グループと，研究代表者のニック・キャンベルおよび国際電気通信基礎技術研究所（ATR）の研究員から成るシステム応用グループによって行われた．本稿では，録音総時間数 1,500 時間におよぶ全コーパスのうち，意味構造グループによって収集されたデータ（以下，神戸データ）について，その内容および研究利用のためのアクセス方法を中心に説明する．また第 3 節では，すでに研究利用されているシステム応用グループによって収集されたデータの概要およびその研究成果についても紹介する．

2．神戸データ

2．1．収録

音声データの収録は，意味構造グループのメンバーによって行われた．話者・収録環境・収録機材・課題などについて，詳細を以下に述べる．

＜話者＞

意味構造グループのメンバーおよびその家族・学内外の友人など，アナウンサーや俳優ではない一般の人たち．

図 1: 収録風景

＜収録環境＞

ほとんどの対話は話者がリラックスして話すことができるよう自宅や大学研究室内で収録された（図 1 参照）．一部の対話は対面式録音ブース（YAMAHA, ANF35S11LL）で収録された．録音ブースでの収録は，2 人の話者が別々のブースに入り，ガラス越しに対話をする形で行われた．録音ブースで収録されたデータは各々の声を相手の声と混ざることなく取り出すことができる．

＜収録機器＞

DAT TCD-T10 （ SONY ） および DAT WALKMAN TCD-D100（SONY）を用いた．

＜課題＞

話者に課題は一切与えられなかった．

＜その他＞

データを収録するにあたり，話者には法律書式での事前承諾を得た．収録後，個人情報の有無など，法律的に問題が生じる可能性のある発話部分には，ブザー音（ビープ音）を施した．メンバー以外の話者には原則として謝金を支払ったが，種々の事情で支払われなかった場合もあった．

2．2．書き起こし

書き起こしには，言語情報の他に「笑い」「咳」「言い淀み」などの非言語情報を含む 31 種類のタグが付与されている．表 1 にタグの全種類を提示する．

表 1: コーパスタグ付け記号

番号	タグの内容	英語名	タグ記号
1	笑い	laughing	@W
2	咳	coughing	<CO>
3	叫び	crying	<CR>
4	ささやき／不明瞭音	murmur/uncertain	<MU>
5	舌鼓	smack	<SM>
6	咳払い	hawk	<HA>
7	吸気音	ingressive	@S
8	呼吸音	breathing	<BG>
9	力み声	creaky	<CK>
10	息混じりの声	breathy	
11	震える声	trill	<TR>
12	音変異	variable	<VA>
13	鼻音化	nasalization	<NA>
14	母音の伸長	lengthening	<LE>
15	焦点	forcus	<FO>
16	流暢でない・言い淀み	disfluency	<DF>
17	ポーズ	pause	＃
18	沈黙	silence	＃
20	上昇	rise	<RI>
21	下降	fall	<FA>
23	始める	start	<-s>
24	終わる	end	<-e>
26	ノイズ	noise	<NO>
27	フィラー	filler	<FI>
28	オノマトペ	onomatopoeia	<ON>
29	繰り返し	repeat	<RE>
30	ビープ音処理		(固有名詞)
31	書き起こし不可		[？]

2. 3. 公開

ここまで紹介してきた音声データは，現在インターネットで公開する作業を進めている．データの公開を可能にするために，文字書き起こしをチェックし，個人情報および問題発言の有無を再確認した．そして該当箇所があった場合には，ブザー音処理を施し，完全に聞こえないようにした．話者に連絡をとり，双方から承諾を得られた対話だけをネット上に公開することにした．

これらの作業は進行中だが，すでに個人情報等のチェックが済んだデータは http://www.speech-data.jp/taba/kobedata/にアクセスすると，実際に聞くことができる．

＜対話音声の再生＞

この WEB ページでは，対話音声とその書き起こし，また対話を視覚的に示したバーチャートを同時に表示・再生できるフラッシュを使用している．このフラッシュには 1）音声再生表示，2）リスト表示，3）全チャート表示の 3 種類の表示機能があり，切り替えて表示することができる．

まず聞いてみたい音声ファイルをクリックすると，『音声再生表示』の状態で 60 秒分の書き起こしがバーチャートで表示される（図 2 参照）．画面右上の「LIST」と書かれたボタンをクリックすると，『リスト表示』に切り替わる．画面右上の「All View」というボタンをクリックすると，『全チャート表示』に切り替わる．

1） 音声再生表示

『音声再生表示』では，対話音声を聞きながらバーチャートと書き起こしを同時に見ることができる（図 2 参照）．

音声を停止している状態でバーチャート枠内でマウスを動かすとチャート内のマウスの位置に縦線が表示され，その位置に対応した書き起こしがバーチャート下部に表示される．音声再生中は，再生位置に合わせて書き起こしが表示されるようになっている．音声の再生・停止はバーチャートの下に表示されているコントロールパネルを操作することによって行う．

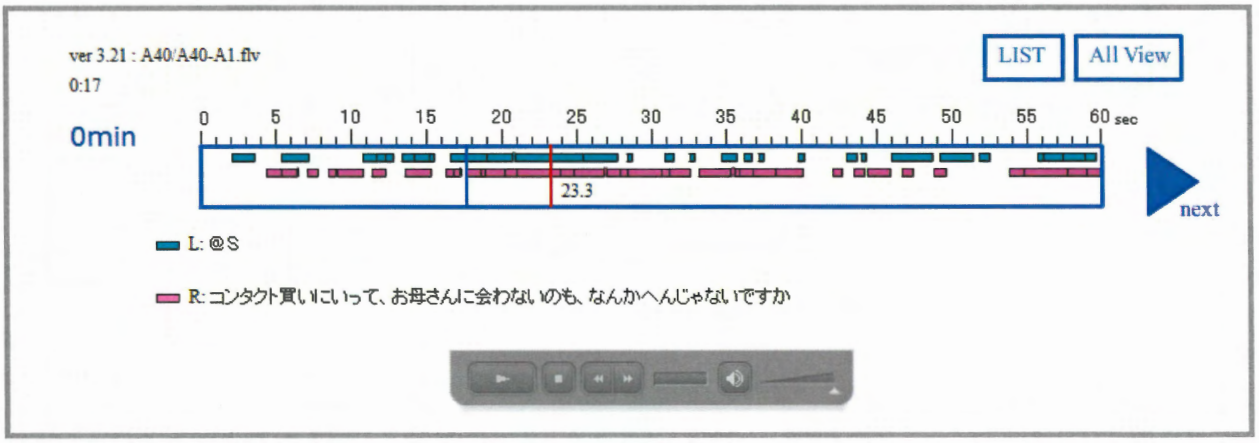

図 2: 音声再生表示の画面

2） リスト表示

『リスト表示』の状態では，書き起こしデータを一覧表で見ることができる（図 3 参照）．表内でマウスを動かしクリックすると，『音声再生表示』の状態に切り替わり，マウスが指している行のスタート時間から音声が再生される．画面右上の『Returen』というボタンをクリックすると，『音声再生表示』に戻る．

ver 3.21 : A40/A40-A1.flv
0:00

speaker	start	end	subtitles
L	2.140	3.599	めがね掛けへんで大丈夫なん
R	4.459	5.405	掛けたらね、＜へへ＞@W
R	5.405	6.460	@S
R	6.460	6.579	(ホ)
L	5.459	7.209	うん、え、それ、だて(じゃ)ないよね(え)
R	7.170	7.829	(違)うんですよ
R	8.579	9.032	今ね
R	9.175	10.859	ニ週間コンタクトが切れてね(え)
L	10.810	11.710	(う)ん、(＜ふふ＞@W)

図 3: リスト再生表示の画面

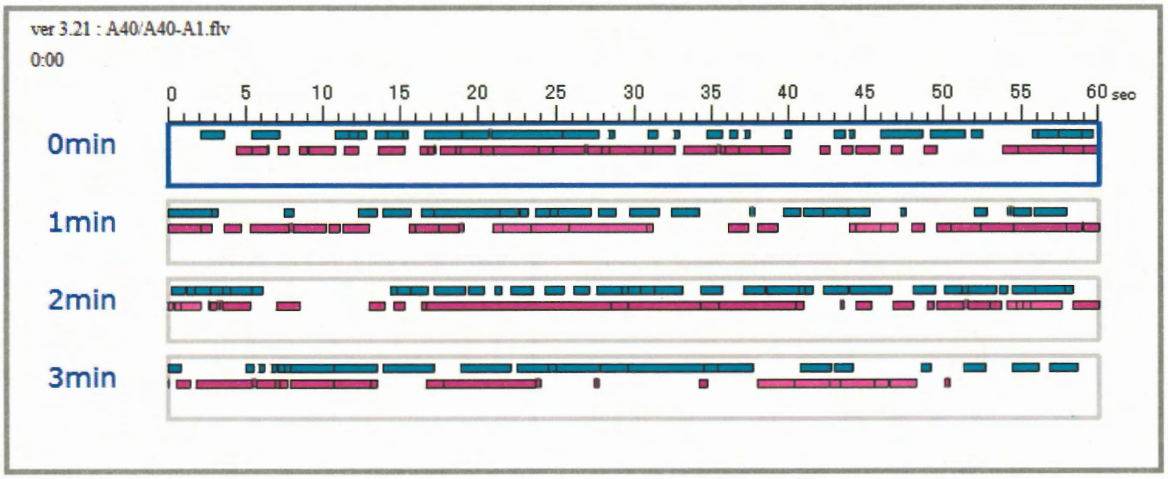

図 4: 全チャート表示の画面

3) 全チャート表示

『全チャート表示』の状態では 1 分一段のバーチャートを複数段一度に表示する（図 4 参照）．マウスを動かすことによって，聞きたい箇所を選択することができ，その状態でクリックすると『音声再生表示』に戻り，選択された段のバーチャートが表示される．

＜語彙検索＞

トップページからは特定の語彙を検索できる「語彙検索ページ」に移動することができる．検索したい語彙を入力し「SEARCH」ボタンをクリックすると（図 5 参照），その語彙を含む発話の一覧が表示される（図 6 参照）．ここでは「ありがとう」と入力し，その検索結果として「ありがとう」を含むデータベース内の全ての発話が表示されている．

「PLAY」ボタンをクリックすると，該当箇所の音声が再生される．コントロールパネルの「巻き戻しボタン」「早送りボタン」を操作することによって，前後の発話を聞くことも可能である．また正規表現を用いて「^ありがとう」とすれば「ありがとう」で始まる発話が，「ありがとう$」とすれば「ありがとう」で終わる発話が，そして「^ありがとう$」入力すれば「ありがとう」のみの発話が表示されるようになっている．

語彙検索をする際に特定の音声ファイルのみにチェックを入れると，そのファイルのみが検索対象となる．図 5 では全てのファイル（A01～A63）にチェックが入っている．

＜その他＞

この WEB ページでは，2 節で説明した書き起こしルールおよび各データの話者・収録環境に関する情報も閲覧できるようになっている．話者については性別・収録当時の年齢・使用方言について書かれている．

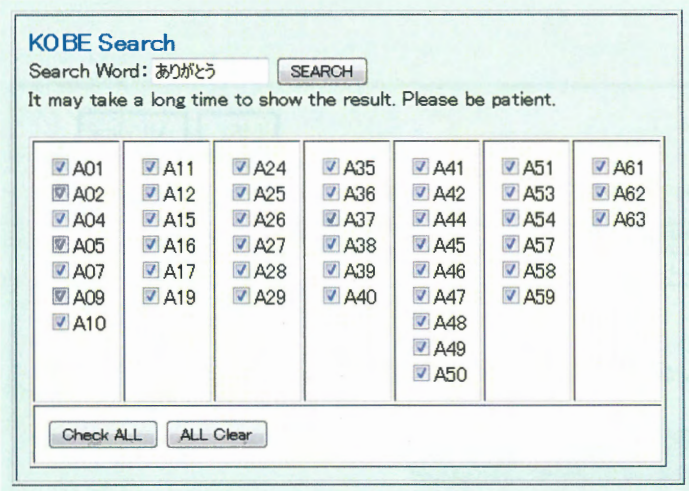

図 5: 語彙検索の画面

start	end	filename	speaker	PLAY	
48.091	48.687	A02-A1	L	PLAY	ありがとう
159.161	160.713	A02-A3	L	PLAY	ありがとうございます
20.767	22.147	A10-A1	R	PLAY	あ、ありがとうございます
403.339	405.544	A12-A0	L	PLAY	ありがとうと思ったけど、(xxxx)だったんですね
1821.510	1823.587	A12-A0	R	PLAY	ありがとう、ありがとうですね、へへへ@w
298.913	300.580	A16-A14	R	PLAY	なるほど、ありがとうございます
180.831	181.823	A16-A15	R	PLAY	ありがとうございます
379.840	381.218	A16-A3	L	PLAY	いや、ありがとうございますって言って
229.126	230.634	A16-A7	R	PLAY	うーん、ありがとう
7.080	8.553	A17-A1	L	PLAY	ありがとうございます
7.487	10.207	A17-A1	R	PLAY	はい、おねがいします、ありがとうござい@wます
43.140	44.634	A19-A1	L	PLAY	はい、あ、ありがとうございました
117.959	118.339	A19-A11	L	PLAY	ありがとう
117.319	118.367	A19-A11	R	PLAY	どうも、ありがとう、ア
235.488	237.801	A25-A10	R	PLAY	一応終わりに、シ、どうもありがとうございました、もう

図 6: 語彙検索の結果を表示する画面

3. ESP コーパスのその他のデータ

ここではデータ整理が済み，すでに研究利用されている ESP コーパスのうち 2 つのサブセットについて，その内容と主な研究成果を紹介する．

3. 1. 日常会話データ

＜データ内容＞

プロジェクト開始時 32 歳であった日本語話者である女性 FAN の 2000 年から 2005 年までの 5 年間に渡る日常会話音声を収録したもの．ヘッドフォン型の小型マイクとミニディスクプレーヤーを用い，話者が可能な限り日常的に機器を装着して収録を行った．この間に妊娠した話者は，2004 年 11 月 11 日の出産当日も平常と変わらず収録を続けた．

全て対話形式で行われたものであるが，対話相手の音声は収録されていない．対話相手は合計 112 人であり，家族・友人・他人・子供に大きく分類することができる．さらに話者による独り言発話も他の話者との対話中に頻繁に行われていたため，それらも収録されている．発話場は自宅や友人宅などであり，会話収録のために実験室や特別な場所に出向くことはなく，いずれも話者が日常生活を送っている場所で収録されたものである．FAN と対話相手は同じ場所で会話していることもあれば，電話で行われた会話を収録したものもある．このデータの総時間数は 600 時間であり，全てが自発的な自然発話である．

＜研究成果＞

1 話者の発話様式が，対話相手に応じていかに変化するかを計量的に示すことに成功した．

図 7 はこのデータの話者が発した 100 時間の対話音声の声の固さと高さを示している．これらの声質の変化は，対話相手に対する親しさや発話の丁寧さと関連していることが分かっている．図 7 左図からは例えば「子供」「他人」と話すときは声が高くそして柔らかくなるのに対して，「家族」「友達」と話すときは低く固くなっていることが分かる．図 7 右図は，左図に示された家族に対する発話を，（家族の）メンバーごとに分析した結果である．ここでも例えば「子供」に話すときは声が柔らかくなるのに対して，「夫」と話すときは固くなるなど，2 つのパラメータの振る舞いから，話者と家族の各メンバーとの人間関係が伺える．詳しくは Campbell&Mokhtari.P.（2003）を参照されたい．

3. 2. 電話対話データ

＜データ内容＞

人材派遣会社を通して選ばれた日本語話者 6 人と非日本語話者 4 人が，週 1 回 30 分ずつ，10 回に渡って行った電話対話を収録したもの．総時間数は 105 時間であり，会話は全て日本語で行われた．収録開始時，10 人の話者は初対面であった． 6 人の日本語話者のうち，男女 1 人ずつ計 2 人の話者は，自らの家族との会話も収録している．

＜研究成果＞

対話相手の違いに応じたものだけでなく，収録回数を重ねるごとに変化する話者間の距離に応じた発話様式の変化を観察することができた．また非日本語話者との会話を収録することにより，相手が日本語話者であるときの会話と比べて，どのように発話様式が変化するかも捉えられている．

ここでは日本語話者である女性 JFA の全収録データに高頻度で出現していた 6 発話（「あぁ（a,a-）」，「あの（ano）」，「でも

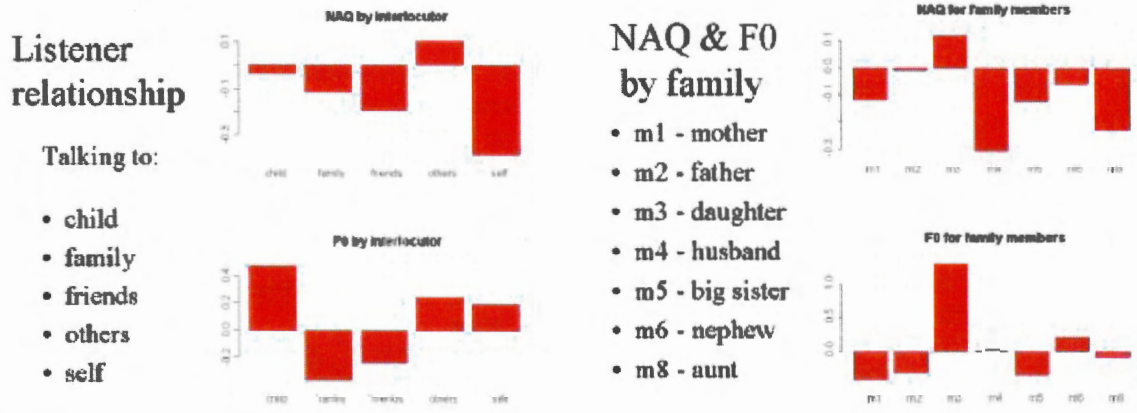

図 7: 対話相手による発話様式の違い
対話相手のグループ（左図）および家族構成員（右図）ごとの NAQ（声の固さ，上図）および F0（声の高さ，下図）．

表 2: 話者 JFA の対話相手に応じて変化する高頻度で現れる発話

JFA:	CFA	CMA	EFA	EMA	JFB	JMA
a,a–	143	145	88	89	138	170
ano	224	277	221	176	209	266
demo	41	24	31	17	89	134
e–	48	51	37	25	74	94
hai	2932	2234	2181	3239	72	33
un,un	1029	546	585	1190	909	1037

（demo）」，「えー（e-）」，「はい（hai）」，「うん、うん（unun）」）を洗い出し，これらの発話の対話相手に応じた使用頻度の違いを，とりわけ日本語話者と非日本語話者を相手にしたときの違いをまとめたものを表 2 に示した．6 人の対話相手のうち，CFA・CMA・EFA・EMA の 4 人は非日本語話者であった．表から，例えば「でも」という発話が非日本語話者との対話データにはあまり多く出現していなかったが，日本語話者との対話データには多く出現していたことが分かる．それとは逆に，「はい」は非日本語話者との対話データには多く出現していたが，日本語話者との対話データにはあまり多く出現していないことが分かる．詳しくは Campbell（2007）を参照されたい．

4. まとめ

本稿では ESP プロジェクトが構築した表現豊かな大規模自然発話音声コーパスのうち，現在公開作業を進めているデータを紹介するとともに，すでに研究利用されている他のデータについてもその内容および研究成果の一部を説明した．これまで自然な音声データの収集と，音響分析に耐え得る高音質のデータの収集は一者択一の課題のように扱われてきた．しかし音声機器の発達や収録のノウハウが確立されるにつれて，このようなジレンマも緩和されつつあるように思われる．ESP プロジェクトでも，3 節で紹介したように実験室的環境ではない環境にて収録された音声データを用いて，音響分析を行うことに成功している．今後，音響分析を必要とする研究においても，自然な環境で収録されたデータの需要が高まることが予想される．

冒頭でも触れたように，音声コーパスを用いた研究は音声工学や言語研究など様々な分野において盛んに行われている．しかし外国語教育，とりわけ日本語教育におけるコーパスの利用は遅れており，その取り組みが今まさに始まったばかりだと言われている（砂川: 2009）．その取り組みとして，日本語学習者が遭遇する様々な場面を設定し，それらに必要な語彙・表現が含まれるようなコーパスをデザインすることは大変重要である．しかし本稿で紹介したような「普通の日本人同士の，普通の会話」を収録したコーパスも，外国語教育に大いに貢献するだろうと思われる．多くの外国語学習者が，母語話者同士の自然な会話を聞き，全く理解できずに愕然とした経験があるだろう．映画やできるだけ自然な会話を題材とした教材も増加しつつあるが，それらが依然として不特定多数の聞き手が理解できるように話されている発話であり，「普通の人の普通の会話」に比べて聞き取りやすく作られていることは否めないだろう．今後外国語教育の分野でも本稿で紹介したような一般の話者による自然な発話コーパスが利用されることが期待される．

［付記］本稿は科学技術振興機構（JST）による戦略的創造研究推進事業（CREST）「表現豊かな発話音声のコンピュータ処理システム」（研究代表者: ニック・キャンベル），日本学術振興会の科学研究費助成金による基礎研究（A）「人物像に応じた音声文法」（課題番号: 19202013, 研究代表者: 定延利之）の成果の一部である．

参考文献

[1] Campbell, N. & Mokhtari, P.: Voice quality; the 4th prosodic dimension, *Proceedings of the 15th International Congress of Phonetic Sciences*, pp.2417-2420, 2003.

[2] Campbell, N.: How speech encodes affect and discourse information: Conversational Gestures, NATO Security through Science, Vol.18, pp.103-114, 2007.

[3] Cowie, R.: Describing the emotional states expressed in speech. *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. 2000.

[4] Erickson, D., Yoshida, K., Mochida, T., & Shibuya, Y.: Acoustic and articulatory analysis of sad Japanese speech, 第 18 回日本音声学会全国大会予稿集, pp. 113-118, 2004.

[5] Iida, A., Campbell, N., Higuchi, F., & Yasumura, M.: A corpus-based speech synthesis system with emotion. Speech communication, Vol.40, No.1, pp.161-187, 2003.

[6] Maekawa, K., Koiso, H., Furui, S., & Isahara, H.: Spontaneous speech corpus of Japanese, Proceedings of LREC 2000, pp.947-952, 2000.

[7] 砂川有里子: コーパスを活用した日本語教育研究, 人工知能学会誌, 24 巻, 5 号, pp.656-664. 2009.

[8] Williams, C. E., & Stevens, K. N.: Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America,* Vol. 52, No.4. pp. 1238-1250, 1972.

[9] http://www.speech-data.jp/taba/kobedata/