TR-IT-0322

研究用自然発話音声データベース解説書 ('99 年度公開版) – 旅行および国際会議申込 –

User's Manual of the Speech Dialogue Database for Spontaneous Speech Recognition (Released 1999) – Travel and Conference Arrangement –

高野 優 Masaru Takano
匂坂 芳典 Yoshinori Sagisaka

1999.12.28

本稿は ATR 音声翻訳通信研究所において構築された研究用自然発話音声データベースの利用解説書である．本データベースは，旅行および国際会議の申し込みに関する日本語の自然発話の模擬会話データであり，自然発話スタイルの文章の発話からなる．本データベースの特徴として，自由発話文の発声に加えて同一話者・同一文章の読み上げ発声が収録されていることが挙げられる．この特徴により，自由発話文と読み上げ文の対照による音声の研究に有用なデータベースとなっている．本稿では，会話音声データベースを利用するための，データベース構成と内容について述べる．

目次

1 はじめに

音声研究における音声データベースの重要性は古くから認識されてきた．これまでも単語や文音声に関するデータベースが作成されてきた [1][2]．一方，連続音声認識技術の発展に伴い，自然発話に関する音声認識への関心も高まっている．このような背景の中で，我々は自然発話データベースの構築を目的として，各種研究目的に共同利用可能な不特定話者を対象とした自然発話音声データベースの構築を進めてきた [3][4][5][6][7]．本稿は，自然発話の特徴の研究に非常に有用と思われる，同一発声内容・二種類の発話スタイル (自由発話および読み上げ) による旅行および国際会議申し込みタスク発話データベースの構造と内容について述べる．2節ではデータベースの構成について，3節では音声データの収録方法の概要について，4節では収録データとそれに関する各種ファイルについて，5節では本データで用いている各種 ID などについて，それぞれ解説する．

2 データベースの構成

本データベースは次の 2 分野の音声データにより構成され，2 枚の CD-ROM に収められている．1 枚目は旅行申し込み会話，2 枚目は国際会議申し込み会話である．

1. 旅行申し込み (CD-ROM1:/TRAV)
2. 国際会議申し込み (CD-ROM2:/CONF)

話者は旅行申し込み会話が 3 名，国際会議申し込み会話が 5 名で，前者は後者に含まれる．

	話者				
	FAK (女性)	FKM (女性)	FKN (女性)	MMY (男性)	MTK (男性)
旅行申し込み対話	○		○		○
国際会議申し込み対話	○	○	○	○	○

3 音声データ収録概要

本データは環境雑音の影響が極めて少ない場所で，DAT へ収録される．発声者はすべてアナウンサー，ナレータ等の職業的話者である．

模擬会話 会話は申込者および担当者の二人による日本語での対話を想定している．ただし，音声データが存在するのは申込者のみである．会話は旅行および国際会議に関する申し込みタスクであり，申込者および担当者による旅行または国際会議の申し込みのための電話を通した会話 (非対面の会話) という設定である．それぞれ以下の手順で録音される．

1. 自由発話

   会話者は事前に基本的な状況設定のみ知らされており，その状況に従って発話が行なわれる．具体的な発話内容および表現は話者に委ねられる．また，自然発話特有の間投詞の挿入，言い淀み，簡単な言い誤り等も話者の創意に任される．

2. 読み上げ

   上記自由発話発声と 1 対 1 で対応している．それぞれ対応する自由発話の発声内容に従った書き起こし文章を作成し，ランダムに並べ直して文脈を正規化したのち，同一話者がそれを読み上げる．

4　データ形式

4.1　模擬会話

模擬会話データは波形データ (WAV)，時刻情報付きローマ字書き起こしデータ (TRS)，境界情報付き音素書き起こしデータ (LBL)，日本語書き起こしデータ (JTEXT)，及び収録情報データ (ENV) より成る．

これらのデータの形式を説明するにあたって必要な概念を，ここで定義する．話者が相手から発話権を受け取って，次にその相手に発話権を渡すまでの単位を**発話**と呼ぶ．また，一つの発話中，長いポーズ (約一秒以上) で区切られる高々 10 秒程度の単位を**発声**と呼ぶ．また，文字に書き起こしたときに，区点「. 」で区切られる単位を**文**と呼ぶ．通常，「発話」「発声」「文」の三つの単位は，一致することが多いが，完全に一致するわけではない．その違いについては，(4.1.4)節で例を示す．

(4.1.1)　波形データ (WAV)

DAT にサンプリング周波数 48kHz(16bit linear PCM) で収録した会話を，16kHz にダウンサンプリングした後に発話単位に切り出したものである．発話単位と発話権の移動とは同じではなく，一回の発話権の間に複数の発話が存在する場合もある．データ・フォーマットは，ヘッダレスの 16bit linear PCM で，big endian で格納されている．

ファイル名の例を以下に示す．

ファイル名フォーマット	話者 ID “_” 発話 ID “_” 発声 ID “.16k”
例	MTK_TF01_001.16k MTK_TF01_002.16k MTK_TF01_003.16k

ファイル名の各フィールドは，話者 ID，発話 ID，発声 ID，波形データの拡張子よりなる．発声 ID は 001 から始まり，会話の発声順に増分 1 で設定される．

(4.1.2)　時刻情報付きローマ字書き起こしデータ (TRS)

各発声をローマ字によって音素単位で書き起こし，音声区間の始端終端に時刻情報を記したものである．1 ファイルは 1 発声からなる．時刻情報は，切り出された発声の始点を 0 として msec 単位で表す．ファイル名は波形データに準ずるが拡張子は “.TRS” である．書き起こしに用いる音素単位は，付録 C の音節表に現れるものとして定義される．

ファイル名の例を以下に示す．

ファイル名フォーマット	話者 ID “_” 発話 ID “_” 発声 ID “.TRS”
例	MTK_TF01_001.TRS MTK_TF01_002.TRS MTK_TF01_003.TRS

図 1に本ファイルの例を示す．“#” で始まる行は，コメントを表し，ファイル中の “-” は，ポーズを表す．

```
295.00    a,a,m,o,sh,i,m,o,sh,i                1145.00
1145.00   -                                    1170.00
1170.00   e,e,zh,e,e,t,i,i,b,i,i,s,a,ng,d,e,s,u,k,a  2790.00
#
#
#
#
#
#
```

図 1: ローマ字書き起こし

(4.1.3) 境界情報付き音素単位書き起こしデータ (LBL)

各発声を音素単位で書き起こし (音素単位に分離不可能な場合は融合表記を用いる), 各音素境界に人手で時刻情報を記したものである. 1 ファイルは 1 発声からなる. 時刻情報は, 該当音素の始点を 0 として msec 単位で表す. ファイル名は波形データに準ずるが拡張子は ".LB" である. 書き起こしに用いる音素単位は, 付録 D の音素表に現れるものとして定義される. 本形式はコメント行を挟んで 5 層にわたって記述されるが, 音声認識に使用されるのは通常第 1 層のみである. 付録 D では第 1 層に関し述べるにとどめる. ファイル名の例を以下に示す.

ファイル名フォーマット	話者 ID "_" 発話 ID "_" 発声 ID ".LB"
例	MTK_TF01_001.LB MTK_TF01_002.LB MTK_TF01_003.LB

図 2に本ファイルの例を示す. "#" で始まる行は, コメントを表し, ファイル中の "pau" は, ポーズを表す.

(4.1.4) 日本語書き起こしデータ

各会話ごとの仮名・漢字を用いた書き起こしである. 文字コードは, ISO-2022 系の日本語拡張 UNIX コード (以下, 日本語 EUC) で書かれている. 1 ファイルは, 1 会話からなる. ファイル名の例を以下に示す. ファイル名は, 会話 ID, 日本語書き起こしデータを示す拡張子 ".JTEXT" より成る.

ファイル名フォーマット	話者 ID "_" 発話 ID ".JTEXT"
例	MTK_TF01.JTEXT MTK_TF01.JTEXT MTK_TF01.JTEXT

図 3に書き起こしの例を示す. この例では, 長すぎる行は折り返して表示してあるが, 実際のデータの一行は, 一発声を表す. そして, 発話権が移動する毎に, 行頭に「担当者」「申込者」という発話ラベルが振られる. また, [...] は間投詞, (...) は言い淀み, 言い誤り, {...} は発話中に挟まれた相づち等 (録音されず) を表す.

```
295.0    a,a   525.0
525.0    m     580.0
580.0    o     640.0
640.0    sh    700.0
700.0    i,m   755.0
755.0    o     815.0
815.0    sh    905.0
905.0    i     1145.0
1145.0   pau   1170.0
         ...
#
         ...
#
         ...
#
         ...
#
         ...
#
```

図 2: 音素単位書き起こし

担当者 ： おはようございます.
交通公社の東京本社内支店でございます.
申込者 ： [ああ] もしもし {はい} [えー] ジェーティービーさんですか.
担当者 ： はい.
申込者 ： [あのーですね] ちょっとお伺いしたいんですけども.
担当者 ： はい.
申込者 ： [えー] いま, 私の持っているパンフレットは, ジャスって書いて
あるんですけども, こちらでいいんですか.
担当者 ： はい, 「ジャスナイスウィング」も, [あの] こちらで取り扱って
おりますので, ご用意させていただくことができます.
申込者 ： [ああ] そうですか.
[えー] こちらは, [あのー] 杉並区の加藤と申しますけども.
担当者 ： はい.
申込者 ： [ちょっとー] 北海道の旅行についてお伺いしたいんですけども,
{はい} よろしいでしょうか.
担当者 ： はい, どんなことでしょうか.
申込者 ： [えーとですねー, えー] 1回だけ, 北海道は行ったんですよね.
担当者 ： はい.
申込者 ： [あの] あちらの, 札幌の方で, ごくごく普通のパックなんですけどね.
[うーん] 今度は, ちょっと遠くの方に行ってみたいと思いまして.
[まあ] というのは, (りゅうふゆ) 流氷ですね, 流氷を見てみたいと
思いまして.
担当者 ： はい.
申込者 ： [えー] やっぱり, それは, 札幌じゃなくて, 網走の方に行かないと
見られないんでしょうね.
担当者 ： はい, [あの] 北から [あの] 東の海岸線に, 流氷が流れ着きますので,
そちらの方でないとご覧いただけません.
申込者 ： [はあはあ] それで, 季節的にはどうなんでしょうね,
[あの] いつがベストなのかしら, [その] 流氷を見るには.
...

図 3: 日本語書き起こし

図3の発声にもあるように，一発声は複数の文からなることがある．また，一発話は，複数の発声からなることがある．例えば，図3の申込者の4回目の発話は2つの発声から成っている．

(4.1.5)　収録情報データ

各会話の収録環境や話者等の情報を示す．文字コードは，日本語EUCで書かれている．ファイル名は会話IDに基づく．ファイル名の例を以下に示す．

ファイル名フォーマット	話者ID "_" 発話ID ".ENV"
例	MTK_TF01.ENV MTK_TF02.ENV MTK_TF03.ENV

図4にファイルの例を示す．各項目はすべて内容が記述されているわけではなく，内容が不明な項目については，値として"NIL"が設定される．

会話ＩＤ:MTK_TF01
収録日時:1991年～1992年
総会話時間:NIL
有効会話時間:NIL
ノイズレベル:NIL
マイクロホン:NIL
ＤＡＴの機種:NIL
ミキサーの機種:NIL
サンプリング周波数:20kHz
発話形態: 自由発話
ドメイン:NIL
トピック:NIL
発声方法: 文発声
対面・非対面: 申込者のみ発話
言語パターン: 日本語－日本語
発呼者: ＭＴＫ__男性__ NIL __東京__ NIL
被呼者: ＭＴＫ__男性__ NIL __東京__ NIL
発呼者発話ラベル: 申込者
被呼者発話ラベル: 担当者 (発話なし)
コメント:NIL

図 4: 収録情報

5 各種記号

5.1 話者 ID

話者 ID はアルファベット 3 文字からなる．先頭の 1 文字は性別 (M は男性，F は女性) を表す．あとの 2 文字は話者特有の ID である．完全に同一の ID であれば，同一話者を表す．

M	TK
Male/Female	話者 ID

5.2 発話 ID

発話 ID は，2 文字のアルファベットに，3 桁の数字が続くフォーマットで表される．アルファベット部分の 1 文字目は “T” または “C” で，“T” は旅行問い合わせ会話を，“C” は国際会議問い合わせ会話を表す．アルファベット部分の 2 文字目は “F” または “R” で，“F” は自由発話を，“R” は自由発話の発話内容の読み上げを表す．数字は発話番号である．

T	F	001
T(旅行)/C(国際会議)	F(自由発話)/R(読み上げ)	発話番号

6 まとめ

研究用不特定話者模擬自由会話音声データベースの構成について述べた．本稿及び本データベースを参照することにより，構築された音声データベースの効率的な共同利用が期待される．

謝辞

作業室の皆様をはじめとする，本データベースの構築にあたり御協力頂いた皆様に感謝いたします．

参考文献

[1] 武田他: “研究用日本語音声データベース利用解説書”, ATR テクニカルレポート, TR-I-28, (1988)

[2] 阿部他: “研究用日本語音声データベース利用解説書 (連続音声データ編)”, ATR テクニカルレポート, TR-I-166, (1990)

[3] Nakamura et. al., “Japanese Speech Databases for Robust Speech Recognition,” Proc. of ICSLP’96, pp. 2199–2202, 1996.

[4] Takezawa et. al., “Speech and Language Databases for SPeech Translation Research in ATR,” Proc. of Oriental COCOSDA Workshop’98, pp. 148–155, 1998.

[5] 塚田他: “研究用自然発話音声データベース解説書 (’97 年度公開版)– 旅行会話タスク –”, ATR テクニカルレポート, TR-IT-0222, 1997

[6] 深田他: “研究用自然発話音声データベース解説書 (’98 年度公開版)– スケジューリング・タスク –”, ATR テクニカルレポート, TR-IT-0279, 1998

[7] 内藤他: “研究用自然発話音声データベース解説書 ('99 年度公開版)– スケジューリング・タスク –”, ATR テクニカルレポート, TR-IT-0303, 1999

付録 A　データベース仕様

収録内容

- 模擬会話 (旅行問い合わせ) 申込者発話
- 模擬会話 (国際会議問い合わせ) 申込者発話

構成データ

データの名称	拡張子	ファイルの単位	内容
波形データ	.16k	発声	16kHz sampling, 16bit linear PCM, big endian, ヘッダレス
ローマ字書き起こしデータ	.TRS	発声	ローマ字による音素単位の書き起こし．音声区間の始終端に時刻 (msec) 情報付与.
音素単位書き起こしデータ	.LB	発声	音素記号による音素単位の書き起こし．音素境界に時刻 (msec) 情報付与.
日本語書き起こしデータ	.JTEXT	会話	仮名漢字混じり文による書き起こし．日本語 EUC.
収録情報データ	.ENV	会話	収録環境，話者情報等. 日本語 EUC.

(註)　「発話」　会話において，話者が相手から発話権を受け取って，次にその相手に発話権を渡すまでの単位.

「発声」　一つの発話中，長いポーズ (約一秒以上) で区切られる高々 10 秒程度の単位.

データ量

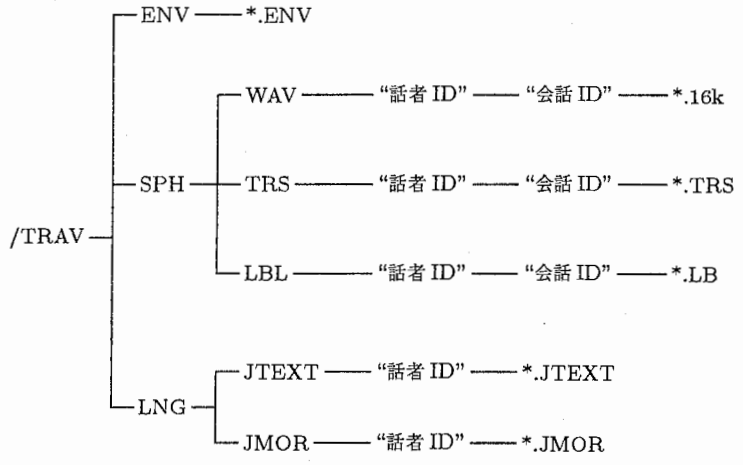
音声データベースのデータ量等に関する情報を示す.
なお，発声時間は音声データから無音区間を除き音声区間のみで計算した.

模擬対話

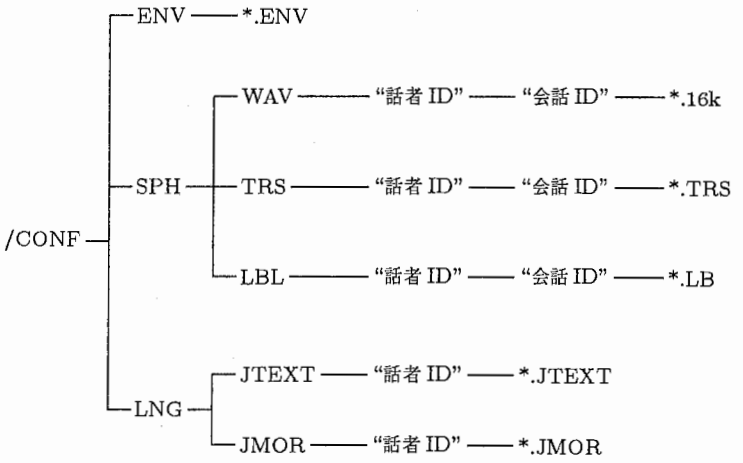
セット名		収録年度	会話数	発声数	話者数 (男性 / 女性)	発声時間	備考
TRAV	自由発話	1991-1992	17	438	(297/414)	0.9h	410MB CD1 枚
	読み上げ	1991-1992	17	363	(297/414)	0.9h	(TRAV 全体)
CONF	自由発話	1991-1992	76	1309	(492/817)	0.8h	350MB CD1 枚
	読み上げ	1991-1992	76	1284	(483/801)	0.9h	(CONF 全体)
total			186	3394	(1381/2390)	3.5h	760MB CD2 枚

付録 B　ディレクトリ構造

CD-ROM1(旅行問い合わせ)



CD-ROM2(国際会議問い合わせ)



付録 C　音節表

時刻情報付きローマ字書き起こしデータにおける，音節表記を以下に示す．

ア	a	イ	i	ウ	u	エ	e	オ	o
カ	k,a	キ	k,i	ク	k,u	ケ	k,e	コ	k,o
サ	s,a	シ	sh,i	ス	s,u	セ	s,e	ソ	s,o
タ	t,a	チ	ch,i	ツ	ts,u	テ	t,e	ト	t,o
ナ	n,a	ニ	n,i	ヌ	n,u	ネ	n,e	ノ	n,o
ハ	h,a	ヒ	h,i	フ	h,u	ヘ	h,e	ホ	h,o
マ	m,a	ミ	m,i	ム	m,u	メ	m,e	モ	m,o
ヤ	j,a			ユ	j,u			ヨ	j,o
ラ	r,a	リ	r,i	ル	r,u	レ	r,e	ロ	r,o
ワ	w,a	ウィ	w,i			ウェ	w,e	ウォ	w,o
ン	ng								
ガ	g,a	ギ	g,i	グ	g,u	ゲ	g,e	ゴ	g,o
ザ	z,a	ジ	zh,i	ズ	z,u	ゼ	z,e	ゾ	z,o
ダ	d,a	ディ	d,i	ドゥ	d,u	デ	d,e	ド	d,o
バ	b,a	ビ	b,i	ブ	b,u	ベ	b,e	ボ	b,o
パ	p,a	ピ	p,i	プ	p,u	ペ	p,e	ポ	p,o
キャ	k,j,a			キュ	k,j,u			キョ	k,j,o
シャ	sh,j,a			シュ	sh,j,u	シェ	sh,e	ショ	sh,j,o
チャ	ch,j,a			チュ	ch,j,u	チェ	ch,e	チョ	ch,j,o
ニャ	n,j,a			ニュ	n,j,u			ニョ	n,j,o
ヒャ	h,j,a			ヒュ	h,j,u			ヒョ	h,j,o
ミャ	m,j,a			ミュ	m,j,u			ミョ	m,j,o
リャ	r,j,a			リュ	r,j,u			リョ	r,j,o
ギャ	g,j,a			ギュ	g,j,u			ギョ	g,j,o
ジャ	zh,j,a			ジュ	zh,j,u	ジェ	zh,e	ジョ	zh,j,o
デャ	d,j,a			デュ	d,j,u			デョ	d,j,o
ビャ	b,j,a			ビュ	b,j,u			ビョ	b,j,o
ピャ	p,j,a			ピュ	p,j,u			ピョ	p,j,o
		ティ	t,i	トゥ	t,u				
ファ	f,a	フィ	f,i			フェ	f,e	フォ	f,o
						ツェ	ts,e	ツォ	ts,o
				テュ	t,j,u				
				フュ	f,j,u				
		ズィ	z,i						

(註)

- ポーズは，“-” で表す．
- 促音は，“q” で表す．
- 長音は，母音を二つ続けることで表す．
- 助詞の「は」「を」は，音に忠実にそれぞれ “w,a”，“o” で表す．

付録 D　音素表

境界情報付き音素単位書き起こしデータにおける，音素表記を以下に示す．ただし，音声認識に重要と思われる第一層のもののみを挙げるにとどめる．

ア	a	イ	i	ウ	u	エ	e	オ	o
カ	k,a	キ	k,i	ク	k,u	ケ	k,e	コ	k,o
サ	s,a	シ	sh,i	ス	s,u	セ	s,e	ソ	s,o
タ	t,a	チ	ch,i	ツ	ts,u	テ	t,e	ト	t,o
ナ	n,a	ニ	n,i	ヌ	n,u	ネ	n,e	ノ	n,o
ハ	h,a	ヒ	h,i	フ	f,u	ヘ	h,e	ホ	h,o
マ	m,a	ミ	m,i	ム	m,u	メ	m,e	モ	m,o
ヤ	y,a			ユ	y,u			ヨ	y,o
ラ	r,a	リ	r,i	ル	r,u	レ	r,e	ロ	r,o
ワ	w,a	ウィ	w,i			ウェ	w,e	ウォ	w,o
ン	N								
ガ	g,a	ギ	g,i	グ	g,u	ゲ	g,e	ゴ	g,o
ザ	z,a	ジ	j,i	ズ	z,u	ゼ	z,e	ゾ	z,o
ダ	d,a	ディ	d,i	ドゥ	d,u	デ	d,e	ド	d,o
バ	b,a	ビ	b,i	ブ	b,u	ベ	b,e	ボ	b,o
パ	p,a	ピ	p,i	プ	p,u	ペ	p,e	ポ	p,o
キャ	ky,a			キュ	ky,u			キョ	ky,o
シャ	sh,a			シュ	sh,u	シェ	sh,e	ショ	sh,o
チャ	ch,a			チュ	ch,u	チェ	ch,e	チョ	ch,o
ニャ	ny,a			ニュ	ny,u			ニョ	ny,o
ヒャ	hy,a			ヒュ	hy,u			ヒョ	hy,o
ミャ	my,a			ミュ	my,u			ミョ	my,o
リャ	ry,a			リュ	ry,u			リョ	ry,o
ギャ	gy,a			ギュ	gy,u			ギョ	gy,o
ジャ	j,a			ジュ	j,u	ジェ	j,e	ジョ	j,o
デャ	dy,a			デュ	dy,u			デョ	dy,o
ビャ	by,a			ビュ	by,u			ビョ	by,o
ピャ	py,a			ピュ	py,u			ピョ	py,o
		ティ	t,i	トゥ	t,u				
ファ	f,a	フィ	f,i			フェ	f,e	フォ	f,o
						ツェ	ts,e	ツォ	ts,o
				テュ	ty,u				
				フュ	fy,u				
		ズィ	z,i						

(註)

- ポーズは，“pau” で表す．
- 促音は，直後の子音表記の先頭１字を重ねて表す．
- 長音は，母音を二つ続けることで表す．

- 助詞の「は」「へ」「を」は，それぞれ “h,a”，“he”，“wo” で表す．
- 雑音記号として，@ls@(唇の音)，@cg@(咳)，@hh@(息)，@lg@(笑い)，#paper_tap#(紙の音)，#key_click#(キーボードの音)，#mug_hits_table#(テーブルの音)，#door_slam#(ドアの開閉音)，#pen_tap#(ペンの音) がある．
- 複数音素の存在が識別できなかった音素間の区切り記号 “,” は省いて表記される．また，境界が決定できなかった場合，複数音素をひとまとめにして境界時刻を付与する．