2008－SLP－74（12）
2008 / 12 / 10

音響的特徴を用いた話し言葉の断片発話単位への分割

瀬戸山勝義 † 柏岡秀紀 †‡§ ニック キャンベル †‡§
† 奈良先端科学技術大学院大学 情報科学研究科
‡ 情報通信研究機構 知識創成コミュニケーション研究センター
§ 国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
E-mail: †{katsuyoshi-s, kashioka, nick}@is.naist.jp

あらまし 現在までの音声合成技術は文を一単位として処理することが多かった．しかし，実対話において，人間は長い発話文を一度に処理することは稀であり，多くの場合，短い断片的な発話を用いる．このような短い断片的な発話を断片発話とし，音声合成の計算処理単位として用いる事を提案する．本稿では,HMM により断片発話の音響的特徴をモデル化し，そのモデルを用いた断片発話単位へのセグメンテーション実験を行なった結果を報告する．実験には，トピックフリーの雑談対話音声を収録した ESP-C コーパスを用いた．

キーワード 断片発話 対話コーパス 話し言葉音声合成 音響的特徴 セグメンテーション

Segmentation of Spoken Language into unit of Utterance Fragment using Acoustics Features

Katsuyoshi SETOYAMA †, **Hideki KASHIOKA** †‡§, and **Nick CAMPBELL** †‡§
† Nara Institute of Science and Technology
†† National Institute of Information and Communications Technology
§ Advanced Telecommunications Research Institute International
E-mail: †{katsuyoshi-s, kashioka, nick}@is.naist.jp

Abstract It is common for speech synthesis technology to process each sentence as one single and independent unit. However, in human speech production, it is perhaps unusual to process a long utterance as a single discrete unit, and typically a series of short utterance fragments is produced in such cases. Such a fragmentary short utterance is assumed to be a minimal discourse unit, and it is proposed here that similar chunks should be used as the basic units for speech synthesis in order to speed-up the calculation processing. In this paper, the acoustic features of such utterance fragments is modeled by HMM, and the paper reports on the result of an experimental the segmentation of a natural speech corpus into optimal units for processing as utterance fragments according to the the model. The ESP-C casual conversation speech corpus was used as material for the experiment.

Key words Utterance fragments, Dialogue Corpus, Spontaneous Speech Synthesis, Acoustics features, Speech Segmentation

1 はじめに

近年, 日常会話に近い対話システム開発のために, 話し言葉音声合成実現に向けた研究が進められている.[1) 2) 3) 4)] その中で問題となっているのが, 断片的な発話やフィラーなどの話し言葉特有の表現である. 現在までの音声合成技術は文を一単位として処理することが多かった. しかし, 実対話において, 人間は長い発話文を一度に処理することは稀であり, ほとんどの場合, 短い断片的な発話を多く用いる. このような短い断片的な発話を断片発話とし, 音声合成の計算処理単位として用いる事を提案する. 対話において断片的な発話は相づちなど, 相手の反応を得やすい発話となっている. よって断片発話を用いることで, 従来の音声合成よりも対話に適した発話音声を生成することが出来ると考えられる.

本研究では, 音響的特徴を用いて統計的手法により, 話し言葉を断片発話単位に分割することを目的とする. 本稿では,HMM により断片発話の音響的特徴をモデル化し, そのモデルを用いて、対話音声を断片発話単位に分割した結果を報告する. 実験には, トピックフリーの雑談対話音声を収録した ESP-C コーパスを用いた.

2 断片発話単位

従来までの音声合成技術は入力テキストとして長い文章を多く用いていた. しかし, 実対話はオーバーラップにより発話が遮られるので, 各発話が長くなることは稀である. よって多くの場合、短い断片的な発話が用いられる. Fig.1 に使用した対話音声の Conversation Chart の一部を示す. これは, それぞれの話者の発話時間が時系列上に色つきのバーで示されているものである.

Fig.1 が示すように, オーバーラップが発話の各所に含まれていることがわかる. オーバーラップを含んでいると言うことはその部分で相づち等を挿入していると言うことであり, 発話の切れ目と考えることができる. また, 実発話でもこのようなオーバーラップを含むことにより, 発話が断片化することは多い.

そこで本研究では, 相づち等の対話相手の反応が挿入する部分を発話の区切りであるとし, 以後その単位を **"断片発話単位"** と定義する. この断片発話単位を用いることで, 対話音声用の話し言葉音声合成が実現できると考えられる.

しかし, 断片発話単位を定めるには, 話者の発声リズムの情報が必要なため, 韻律的な情報が不可欠であると考える. そこで本研究は音響的特徴を用いて話し言葉を断片発話単位に分割することを目的とする.

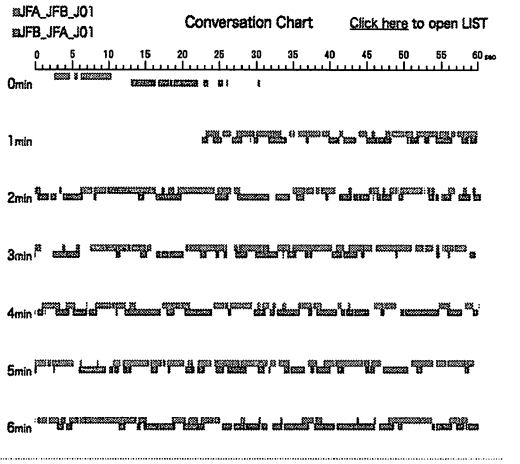

Fig.1 Conversation Chart

3 ESP_C コーパス

本稿では音声資料として ESP_C コーパスを用いたので, そのコーパスについて説明する. **ESP_C コーパス**[5)] は, JST/CREST の「表現豊かな発話音声のコンピュータ処理システム」プロジェクトによって作成された, JST/ATR Expressive Speech Processing Corpora のサブセットである. このコーパスは 2 名の話者による電話での実対話の録音と書き起こしのセットで構成され, 話者は合計 10 名 (男女ともに 5 名ずつ, 日本語を母語としない者 4 名を含む) である. 電話による 1 セッション 30 分の対話を面識のない状態から, 週 1 回収録を行い, 計 10 回 (日本語を母語としない者を含む組み合わせでは 5 回) 収録している. 本稿では, 日本人女性 2 名と日本人男性 1 名の計 3 名を対象とした. 図 2 に ESP_C の書き起こし例を示す. 書き起こしには, 発話者, 発話開始時間, 発話時間が含まれている.

ESP_C コーパスの特徴として, タスクのない全くの雑談であることが挙げられる. 雑談であるということは, より対話的であるので発話毎のオーバーラップが多いと考えられる. そのため, 各発話が朗読調の音声と比べても短い単位になると考えられる. 実際に ESP_C コーパスの書き起こしデータを

JFA_JMA_E04 596.440 1.518 いやーんでも
もうね
JFA_JMA_E04 598.023 0.947 あのー
JFA_JMA_E04 598.994 1.427 一回やってる
ことだから
JMA_JFA_E04 600.335 0.446 うん
JFA_JMA_E04 600.782 0.499 うん
JMA_JFA_E04 601.070 0.440 まー
JMA_JFA_E04 601.543 0.590 ほんで
JFA_JMA_E04 601.598 0.398 また
JMA_JFA_E04 602.189 0.800 そんな
JMA_JFA_E04 603.245 1.133 あれもないし

Fig.2 ESP_C コーパスの例

見ても, 朗読調のコーパスよりも短い単位の発話が多く見受けられた. これより,ESP_C コーパスが話し言葉音声のデータとして妥当でありかつ, 断片発話単位のデータとして使用できると考えられる.

4 断片発話単位への分割手法

断片発話をモデル化するために, ポーズを用いた手法と韻律情報を用いた手法を比較する.

4.1 パワーを用いた発話区間検出

ポーズと韻律情報を用いた断片発話抽出を行うために前処理として, パワーを用いて発話区間の抽出を行なった. 手法は, 音声データからパワーを抽出し, その値が閾値より上だった場合, 発話区間として抽出するものである. Fig.3 にパワーを用いて抽出した発話区間に対し, ESP_C コーパスの書き起こしに記述されている, 発話開始時間と発話時間長を用いて, コーパスの書き起こし発話との対応付けを行なった結果の一例を示す. 図形内に含まれている数字はコーパスにおける発話番号を示している. この数字が同じものはコーパス中では同じ発話である. なお, 図形外の数字は時間を示している.

4.2 ポーズを用いた断片発話単位への分割

ポーズを用いた断片発話抽出を行なった. 手法は, 抽出した発話区間と次の発話区間の間の時間長をポーズとし, これがある閾値以上ならそこを発話の境界とするものである. Fig.3 から分かるように, たとえポーズが短くても別発話となるものはあるし, 逆にポーズが長かったとしても, 同一発話となることもある. よって, ポーズによる閾値処理だけで断片発話を抽出することは難しいと考えられる. 今回の実験では 100msec を閾値として設定した.

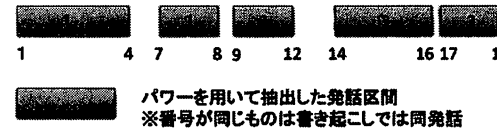

Fig.3 パワーを用いて抽出した発話区間とコーパス書き起こし発話との対応付け

4.3 韻律特徴を用いた断片発話単位への分割

断片発話をモデル化する上でピッチパターンの情報は不可欠である. このため, 断片発話の抽出には, ピッチパターンを考慮に入れたモデルを作成しなければならない. そこで今回は HMM[6] を用いて断片発話抽出のモデルを作成した. なお, 本研究では, ピッチパターンを付与した HMM を構築するために, HMM-based Speech Synthesis System[7] の HMM モデルを用いた. これは, 音声認識等に用いられる離散分布 HMM や, 連続分布 HMM を適用するだけではピッチパターンを表現する事は出来ないためである.

断片発話のモデル化の学習データ作成のために, 以下の手順でパワーを用いて抽出した発話区間に対してラベリングを施した.

1. パワーを用いて抽出した発話区間の発話開始時間と終了時間, コーパス中の発話開始時間と発話時間長を用いて, パワーを用いて抽出した発話区間とコーパス書き起こし発話との対応付けを行う. この結果の一例を Fig.3 に示す.
2. 1 のデータに対して, **"連結"**, **"無音"** , **"独立"**, **"無声音"** のラベルを付与する. 方法は, 現在の発話が次の発話と書き起こし行で同じ発話であるならば **"連結"** , 発話と発話の間は **"無音"** , 書き起こしの発話に対しパワーにより発話が一つしか抽出できなかった場合 **"独立"** , 連結ラベルと連結ラベルの間には **"無声音"** である. Fig.3 に対して, ラベリングした結果を Table 1 に示す.

このようにラベリングされたデータを, 断片発話抽出のための学習データとした.

Table1　Fig. 3 に対するラベリング結果

発話開始時間	発話終了時間	ラベリング結果
1	4	連結
4	7	無声音
7	8	連結
8	9	無音
9	12	独立
12	14	無音
14	16	連結
16	17	無声音
17	19	連結

5　韻律情報を用いた断片発話単位への分割実験

5.1　実験条件

4 章で説明した手法を用いて，断片発話分割実験を行なった．使用した特徴は 12 次元 MFCC, δMFCC, $\delta\delta$MFCC, logf0, δlogf0 である．学習データとして，JFA_JFB_J01〜J05(30 分 ×5 セット) の音声，評価データとして，JFA_JFB_J06(30 分× 1 セット) を用いた．

5.2　結果と考察

実験結果と考察を行う．Table 2 にポーズのみで分割した断片発話数と，コーパス中の断片発話数と，HMM を用いて分割した断片発話数を示す．この数が多いほど，発話を細かくセグメントしており，細かい発話単位に分割したことになる．なお，コーパス中の断片発話数は書き起こし行の数とした．結果から分かるように，HMM を用いて断片発話単位に分割した方が，より細かい断片発話単位に分割できる事が分かる．よって，HMM を用いた手法の方が，ポーズを用いた手法よりも発話を細かく分割している．

Table2　各手法における抽出断片発話数

	断片発話数
ポーズによる抽出	525
ラベラによる抽出	781
HMM による抽出	1055

Table 2 より，HMM を用いた手法の方が，コーパス中の断片発話数よりも多くの単位を抽出したことが分かる．この結果について細かく分析するために，HMM を用いて抽出した各ラベル数とラベラにより抽出した各ラベル数の比較を Table 3 に示す．

Table3　HMM とラベラから抽出した各ラベル数の比較

	HMM の抽出ラベル数	ラベラからの抽出ラベル数
連結	2773	1735
独立	928	308
無音	1698	736
無声音	1855	1316

結果をみると，HMM を用いた方が多くのラベルを抽出していることが分かる．これは，HMM が“独立”として分割した発話が，元のコーパスの発話を細分化したことが原因である．しかし，HMM が“独立”として分割した断片発話を聴取してみると，ほとんどが「うん」や「あのー」等の感動詞やフィラーであったので，HMM が誤った細分化を行なったとは言えないと考える．また，コーパス中に複数存在した，長い断片発話も HMM を用いることでさらに細かく分割することが出来た．よって，断片発話分割の手法として HMM を用いた手法は妥当であると言える．

また,HMM を用いて分割した音声を聴取してみたところ, 言語的な部分で分割されていることが確認された．そこで実際にどの程度言語的部分で切れていたのかを確認した．その手法について説明する．

1. Julius[8] の音素セグメンテーションキットを用いて, コーパスの対話音声に対して音素セグメンテーションを行い, 音素単位でのアラインメントを取得する．
2. HMM より得た, 対話音声の断片発話単位分割境界に関する時間情報と, 先ほど得た音素セグメントの時間情報との対応を取る．この操作をすることにより,HMM での分割により得られた, 言語情報がわかる．
3. HMM で分割した発話の分割境界が,Fig.4 のように元の書き起こしの形態素解析境界とどれだけ一致するのかを調べる．この例で説明すると,「今日は」の部分は形態素の境界と一致してい

るので正解とするが,「天」「気」に関しては境界と一致していないので不正解となる.

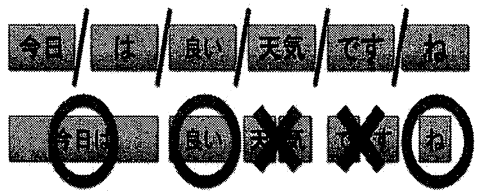

Fig.4 言語境界一致率の計算

上記の方法で計算してた HMM による言語境界との一致率の結果を Table4 に示す.

Table4 HMM によって分割した断片発話単位の言語境界との一致率

一致数	総断片発話数	一致率
769	1,083	71%

この結果より,HMM を用いた断片発話単位分割の手法が誤った言語境界で分割を行なっていないと考えることが出来る.

6 評価

各断片発話単位分割手法に対して, 評価を行う. ここでは評価データに JFA_JFB_J06, JFB_JFA_J06, JMA_JFA_E01 を用いた. 各話者の学習は,JFA_JFB_J01〜J05, JFB_JFA_J01〜J05, JMA_JFA_C01〜C05 をそれぞれ用いた.

次に評価方法について説明する. 断片発話単位は相手の発話が挿入する部分を発話の区切りとするので, 分割された発話が次発話に重なっていないかで, その発話が適切に断片発話単位に分割されたのか判定する. 例えば,Fig.5 のような分割が行われた場合を考える.

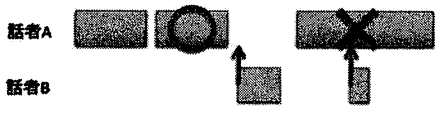

Fig.5 分割発話例

話者 A の二つ目の発話は話者 B の発話開始時間よりも先に終了しているので, 適切に分割できたと考えられる. 話者 A の 3 つめの発話の途中には話者 B の発話があるので, 発話の分割点があると考えることができる. しかしこの例の場合分割されていないので, この発話に対しては, 不適切な分割をしていると考える. このように, ある発話時間内に, 別話者の次発話の開始時間を含むか否かで, 断片発話単位として適切に分割できたかどうか評価する.

しかしこの評価方法の場合, 相手話者が相づちなどの発話をした場合に起こるオーバーラップ時間について考慮されていない. そこで先ほど説明した評価方法を拡張するために, どれだけのオーバーラップを許容して良いか考える. まず, コーパスの書き起こし発話に対して, 各発話のオーバーラップしている時間長を調べた. その結果を多くのオーバーラップ時間長は 0〜1sec の間に存在することが分かった. そこで, オーバーラップ時間長が 0sec〜1sec の間のオーバーラップ時間長の発話数を調べた. その結果を Fig.6 に示す.

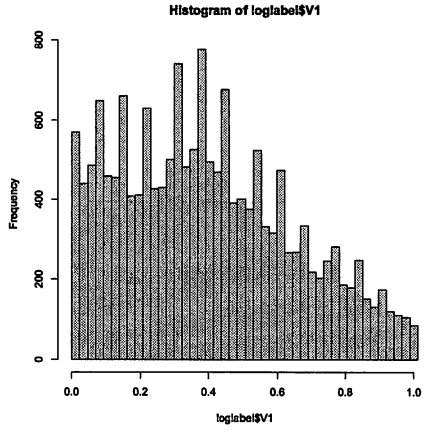

Fig.6 各発話に対するオーバーラップ時間長 (0〜1sec)

Fig.6 の結果を見ると, オーバーラップの時間は 500msec 内に多く集中していることが分かる. また, オーバーラップ発話によく出現する相づちに関して, その発話時間を計測したところ平均 500msec であった. 今回オーバーラップ時間はコーパスの書き起こし単位で計算したが, 書き起こし単位の分割が断片発話単位として正しいとは限らない. そこで,

本稿では実発話のオーバーラップで良く発生する相づちの平均発話時間 500msec をオーバーラップの許容範囲時間として採用することにした．つまり，ある発話に対して，別話者の次発話の発話開始時間が，発話時間内に存在しても，その発話開始時間とが発話の終了時間との差が 500msec 内にあるなら，その発話は断片発話単位に適切に分割できていると考えるということである．

以上の点に注意して，各断片発話分割手法に対して，断片発話単位分割率を計算した．断片発話単位分割率は以下の式を用いた．

$$\text{断片発話単位分割率}\ (P) = \frac{\text{分割に成功した発話数}}{\text{全発話数}} \quad (1)$$

この式を用いて計算した各話者における断片発話分割率の結果を Table5 に示す．ここでのラベラ単位とはコーパスの書き起こしである．

Table5　各発話単位における各話者での断片発話単位分割率

	ポーズ単位	ラベラ単位	HMM単位
JFA	39%	43%	42%
JFB	23%	35%	56%
JMA	39%	49%	64%

Table5 の結果より,HMM を用いて断片発話単位に分割した方が，ラベラ単位と比べても適した分割が行えることが分かる．JFA に関しては HMM よりもラベラの方が適した分割を行なっている．これは,JFA が他の話者に比べて発話時間が長く，その発話を HMM で細かく分割したのが原因である．HMM での分割単位とポーズでの分割単位と比較してみても,HMM を用いた方が断片発話単位に適した分割を行えていることを確認出来る．

7　おわりに

話し言葉を断片発話単位に分割するために，ポーズを用いる手法と韻律情報を用いる手法を比較した．その結果，韻律を用いる手法の方が断片発話単位に分割する上で妥当であることが分かった．

今後の課題は，分割した断片発話単位を音声合成に適用するために，断片発話単位の韻律特徴を明らかにしていくことが挙げられる．発話断片の韻律特徴に関して小磯ら [9] が韻律のクラスを“終了”，“継続”に分類して研究を行なっている．しかし，実発話において韻律クラスは“終了”，“継続”だけでなく，発話者がその発話を終了したのか継続しているのかが分からない“曖昧”の 3 種類の韻律クラスに分けられると考えられる [10]．

また，榎本ら [11] が対話音声に良く出現するオーバーラップ発話の発話様式に関して“情報提供”，“あいづち”，“情報要求”等の 5 分類に分けて研究を行なっている．そこで断片発話単位の韻律特徴に関してこれらの先行研究を参考にし，どのように分類できるのかを検討する必要がある．

参考文献

1) 伊藤芳幸, 岩野公司, 古井貞熙, “話し言葉音声合成における韻律制御要因の有効性の評価”, 日本音響学会 2008 年秋期研究発表会 論文集,1-4-16,pp.271-272.
2) 中村匡伸, 岩野公司, 古井貞熙, “話し言葉音声の音響的言語的特徴の分析”, 電子情報通信学会技術研究報告, Vol.106, No.78,pp. 19-24,SP2006-4.
3) 林由紀子, 松原茂樹, “自然な読み上げ音声出力のための書き言葉から話し言葉へのテキスト変換”, 情報処理学会研究報告,Vol.2007,No.47,pp. 49-54,2007-NL-179-9.
4) 広瀬啓吉, 阪田真弓, “対話音声と朗読音声の韻律的特徴の比較”, 電子情報通信学会論文誌, Vol.J79-D-2, No.12,pp. 2154-2162.
5) Feature Extraction and Analysis for Speech Technology, http://feast.atr.jp/
6) The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/.
7) The HMM-based Speech Synthesis System, http://hts.ics.nitech.ac.jp/.
8) 大語彙連続音声認識エンジン Julius, http://julius.sourceforge.jp/.
9) 小磯花絵, 堀内靖雄, 土屋俊, 市川熹, “ 先行発話断片の終端部分に存在する次発話者に関する言語的・韻律的要素について ”, 電子情報通信学会技術報告,NLC95-72,pp.25-30.
10) 瀬戸山勝義, 柏岡秀紀, ニック キャンベル, “断片発話の韻律特徴による分類”, 日本音響学会 2008 春期研究発表会 論文集, 3-Q-19, pp415-416.
11) 榎本美香, 土屋俊, “オーバーラップ発話の評定方法とその基礎統計:日本語地図課題対話に対して”, 情報処理学会研究報告, Vol.99, No.108, pp. 145-150, 99-SLP-29-25.