社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音節波形接続型音声合成における自動セグメンテーションの影響

藤尾　聡†　村上　仁一†　池原　悟†

† 鳥取大学工学部，鳥取県

あらまし　音節波形接続型音声合成は，波形編集型の音声合成方式の 1 種で，音響的なパラメータを使用しないで，言語的なパラメータのみで音声合成を作成することを特徴としている．この論文では，この音節波形接続型音声合成において，自動セグメンテーションによる影響を報告する．

キーワード　音声合成，波形編集方式，自動セグメンテーション，言語情報

Influence of automatic segmentation for Concatenating Syllabic Components based on Positional Features

Satoshi FUJIO†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Faculty of Engineering,Tottori University

Abstract　The "Concatenating Syllabic Components based on Positional Features and Mora Information and Accent (CSCMA)" is a kind of corpus-based concatenative speech synthesis method. The features of CSCPMA is that only the language information is used to select the wave form. And we have been applied the CSCPMA to phrase speech synthesis and obtained good results. However, one weak point this method is that this method needed a hand labeled data. It's too cost. So we try to use the automatic labeling data used HMM. The result of experiments show that the difference between hand labeling and automatic labeling is little.

Key words　Speech Synthesis, Corpus Based , Automatic Segmentation, Language Model

1. ま え が き

音節波形接続型音声合成 [2] は，言語的なパラメータのみで音声合成を作成する方法である．これは，信号処理を行わずに接続することにより，自然性の高い合成音声が作成できる．この方式は，過去に固有名詞 [3]，普通名詞 [4]，そして文節 [6] を対象として，音声合成が行われた．その結果，品質の高い合成音声が得られた．

しかし，音声合成を作成する場合に必要となるラベリングデータは，人手によって作成されるため，コストがかかる．そこで，本研究では文節発声の音声に対して，不特定話者の自動ラベリングをおこない，音節波形接続型で作成した合成音声の品質を評価する．

なお，自動ラベリングの研究は，従来から多くの研究機関で行われており，HMM 方法とベイズ確率を用いた統計的・確率的モデルによる方法 [10]，ルールベースを用いる手法 [11]，知識処理に基づく方法 [12] などが報告されている．

2. 音節波形接続方式

2.1 音声合成に使用する素片

音節波形接続型音声合成は，波形編集型の音声合成方式の 1 種で，音響的なパラメータを使用しないで，言語的なパラメータのみで音声合成を作成することを特徴としている．具体的には，音節波形接続型音声合成では，表 1 の条件の一致する音節素片を接続して，音声を合成する．

表 1　実験に用いた音節素片

中心の音節
直前の音素 (前音素環境)
直後の音素 (後音素環境)
文節中のモーラ位置
文節のモーラ数
文節のアクセント型

音節波形接続型音声合成の例として「むかって (mu/ka-q/te)」を音声合成する場合を以下に示す．なお,「_」は音の強弱（アクセント）を表している．() 内強調部は，実際に選択される部分を示している．

むかって (mu_/ka-q/te)
＝昔の (**mu_**/ka/sh**i**/no)
＋使った (ts**u**_/**ka-q**/ta)
＋歌って (u_/ta-q/**te**)

2.2 韻律，継続時間，調音結合の情報

一般に音声合成を行う場合，韻律の扱いが重要である．CHATR などの通常の波形接続型音声合成では，ToBI モデルや藤崎モデルで韻律モデルを推定し，推定したピッチ周波数に類似した音素素片を録音した音声データのなかから選択する．しかし，特定話者の単語発話を合成音声に使用する場合，単語のモーラ情報 (モーラ数とモーラ位置) が決まれば，単語によらずピッチ周波数がほぼ決定されることが知られている [3]．また，一般名詞の場合，名詞のモーラ情報に，名詞のアクセント型を加えることによって，非常に高い品質の音声が得られる [4]．文節発声の場合では，発話速度が遅い音声の場合には，文節単位でゆっくりと区切るためピッチが初期化される．それにより，文節発声の音声も一般名詞のみの発話と同様に扱うことができる [6]．

音節波形接続型音声合成は，これらのことがらを利用して，韻律情報を，主に，モーラ数，モーラ位置，アクセント位置から得ている．また，音節継続時間の情報は，主に，音節の前後環境と，モーラ長およびモーラ位置から得ている．そして，調音結合の情報は，主に，音節の前後の環境から得ている．

2.3 波形接続に関する補則

音節波形接続型音声合成における波形接続に関する補則を以下に述べる．

2.3.1 連 続 母 音

音節波形接続方式で作成された合成音声は，音声素片の接続部に違和感を生じる．特に違和感を生じる部分は，母音や撥音，促音が連続する部分である．これらの音節は前後の音が連続的に変化する部分であり，音節境界がはっきりせず，切り分けるのは困難である．これは，母音や撥音，促音が連続する場合，連続母音として扱うことにより，違和感を軽減できる [6]．

2.3.2 音量，発話速度

音節波形接続方式では信号処理を加えないため，合成音声に使用する素片の音量の差が，音声の品質に直接影響する．そこで，音声の録音した時間帯が分かっている場合では，録音した時間帯が近い素片を選択し，音量や発話速度の均一化を行う [4]．

2.3.3 音節素片の接続部

波形接続型音声合成では，接続部の違和感の発生が音声の自然性に大きく影響する．そのため，接続部における 2 素片間の波形の位相を考慮し，接続部の振幅の差がゼロに近づくように接続する．具体的には，あらかじめラベル付けされた素片開始時間と素片終了時間をもとに，振幅が負から正に変わる部分を，波形が短くなる方向 (開始時間は進む方向，終了時間は戻る方向) に探し，抽出する位置を修正する [2]．現在は，まだこの部分は人手による作業に依存している．

3. 自動音節ラベリング

本研究では，自動音節ラベリングを行うために HMM を用いる．具体的なツールとして，HTK [9] を使用する．自動音節ラベリングの手順を以下に示す．

（1） 既に人手でラベリングされた大量の音声データを，Viterbi アルゴリズムを用いて，音節 HMM の初期モデルを作成する．

（2） 音節 HMM の初期モデルを，Baum-Welch アルゴリズムを用いて，音節 HMM 初期モデルを再推定する．

（3） 作成された音節 HMM を用いて，音声合成に用いる音声データベースに対し，Viterbi アルゴリズムを用いて，音節境界位置を計算する．

4. 評 価 実 験

4.1 音声合成に用いる音声データベース

合成音声の対象として，複数の電子辞書から重文複文を抽出した日英対訳の例文集の文を使用する．この例文集は機械翻訳を目的にしたものである．この例文集に収録されている 1000 文を使用し，女性話者 (プロのナレーター) に，文節発声で遅く発声した音声を音声データベースとして用いる．この収録された音声データベースに対して自動ラベリングを行う．収録した音声発話の一部を表 2 に示す．なお，表中の”-”は文節の区切りであり，収録時にポーズを入れて収録した．

表 2 収録した音声発話の一部

番号	文例
0001	これは-人々に-愛唱されている-古い-民謡の-一つです
0002	この背広に-合いそうな-ネクタイを-何本か-見せてください
0003	投手は-次の-打者に-セカンドフライを-打たせて-アウトにした

4.2 音声合成を行うデータ

評価実験のときに合成音声と自然音声と比較するために，音声データベース [4.1] 章に同一の発話内容が存在する音声を合成する．評価実験に使用する 100 文節の，モーラごとの文節数の割合を表 3 に示す．

表 3 実験に使用する 100 文節中のモーラごとの内訳

モーラ数	文節数
4mora	17
5mora	70
6mora	13

これらの発話内容に対し，自動ラベルを用いた場合と手動ラベルを用いた場合の音声合成を作成する．作成した音声の例を表 4 に示す．

表 4 作成した音声発話の一部

番号	作成音声
1	本性を
2	政界を
3	一生に

4.3 自動ラベリングのための学習データ

学習データとして，ATR の単語発話データベース Aset に収録されている，奇数番号データの女性話者 10 名 (1 話者につき 2,620 単語)，合計 26,200 単語を使用する．なお，自動ラベリングを行う音声合成に用いる音声データベース [4.1 章] の話者は，学習データに含まれないため，不特定話者の自動ラベリングになる．

4.4 自動ラベリング

自動ラベリングの手法として，HMM に基づく Vitebi アルゴリズムを利用する．具体的には HTK [9] を利用する．特徴パラメータとして MFCC を，共分散行列には Diagonal-covariance を使用する．その他の実験条件は表 5 に示す．

表 5　自動ラベリングに用いたパラメータ

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態 (連続分布型)
stream 数	3
特徴	12 次 MFCC+Δ 12 次 MFCC
パラメータ	+対数パワー+Δ対数パワー (計 26 次)
連続型 HMM の 初期モデル 混合分布数	(母音・撥音・無音・連続母音) MFCC 10, Δ MFCC 10, 対数パワー 2, Δ対数パワー 2 (その他の音節・ue・uq) MFCC 4, Δ MFCC 4, 対数パワー 1, Δ対数パワー 1

4.5 音声合成の評価方法

音声合成の評価は，オピニオン評価，対比較実験の 2 種類で行う．非試験者は音声研究に関わったことのない学生 5 名とする．

（1） オピニオン評価

音声の自然性を調べるために，オピニオン評価を行う．自然に聞こえた度合を 5 段階 (1 が最も不自然，5 が最も自然) で評価する．

（2） 対比較実験

作成した音声の評価のために，対比較実験を行う．対比較実験は，自然音声，手動ラベルを用いた合成音声，自動ラベルを用いた合成音声の 3 種類を使用し，以下の 3 回行う．

A　自然音声-手動ラベル

B　自然音声-自動ラベル

C　手動ラベル-自動ラベル

各組合せにおいて，同じ内容の文節発声の 2 種類の音声を連続して聴き，どちらの音声が自然に聞こえたかを判定する．

5. 実 験 結 果

5.1 自動ラベルと手動ラベルの音節境界位置の差

合成音声「学校に (ga-q/ko-u/ni) を作成したときの，自動ラベルと手動ラベルの音節境界位置の差の例を表 6 に示す．

学校に (ga-q/ko-u/ni)

＝学校を (**ga-q**/ko-u/o)

＋実行に (ji-q/**ko-u**/ni)

＋本当に (ho-N/to-u/**ni**)

表 6　自動ラベルと手動ラベルの音節境界位置の差

音声	用いた音節	自動ラベル (ms)	手動ラベル (ms)
学校を (ga-q/ko-u/o)	ga-q	100〜500	100〜507
実行に (ji-q/ko-u/ni)	ko-u	470〜840	502〜868
本当に (ho-N/to-u/ni)	ni	840〜1130	856〜1122

この例では，自動ラベルと手動ラベルに差がないことが分かる．オピニオンスコアでは，手動ラベルを用いた合成音声では 3.8，自動ラベルを用いた合成音声では 3.4 であった．

5.2 自動ラベルと手動ラベルの音節境界位置の平均の差

5.2.1 音節境界位置の平均値と標準偏差

合成音声 100 文節の音節境界位置の自動ラベルと手動ラベルの差の平均値と標準偏差を表 7 に示す．

表 7　音節開始時間，音節終了時間および音節継続時間の自動ラベルと手動ラベルの差

	平均値 (ms)	標準偏差 (ms)
音節開始時間	-26.6	42.0
音節終了時間	-22.6	43.8
音節継続時間	3.9	62.0

表 7 から，自動ラベルと手動ラベルにおける，音節開始時間の差の平均値は-26.6ms，音節終了時間の平均値の差は-22.6ms となり，全体的に，音節開始時間，音節終了時間ともに，自動ラベルデータは手動ラベルデータに比べ，音節境界位置が早くなることが分かる．

5.2.2 音節境界位置の分布図

自動ラベルと手動ラベルの音節開始時間の差の分布図を図 1 に示す．

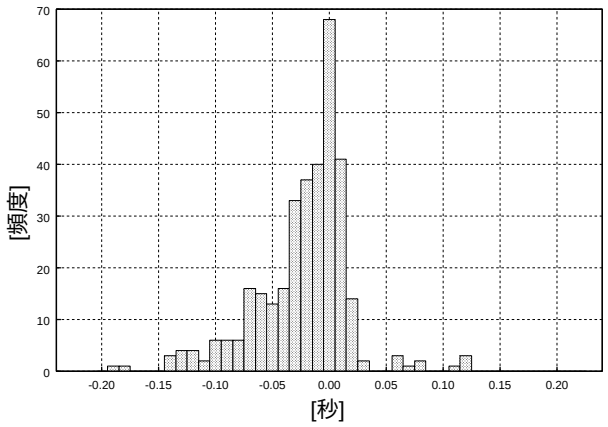

図 1　音節開始時間の差の分布図 (自動ラベル-手動ラベル)

図 1 から，全体的に，自動ラベルと手動ラベルの音節開始時間は，早めにラベリングする傾向があることが分かる．また，音節開始時間の差の頻度として，最大なものは，-0.05〜0.05 の範囲であることが分かる．

図 1 において，自動ラベルと手動ラベルで大きく異なったデータを以下に示す．

「歩いていた」「ta」-199ms

「肝臓を」「o」117ms

自動ラベルと手動ラベルの音節終了時間の差の分布図を図 2 に示す．

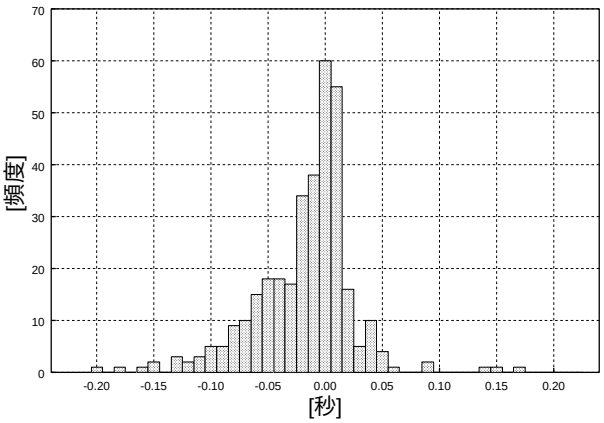

図 2　音節終了時間の差の分布図 (自動ラベル-手動ラベル)

図 2 から，全体的に，自動ラベルと手動ラベルの音節終了時間は，早めにラベリングする傾向があることが分かる．しかし，音節終了時間の差の頻度として，最大なものは，-0.05〜0.05 の範囲であることが分かる．

図 2 において，自動ラベルと手動ラベルで大きく異なったデータを以下に示す．

「結婚する」「ke-q」-201ms

「復興を」「ko-u」168ms

自動ラベルと手動ラベルの音節継続時間の差の分布図を図 3 に示す．

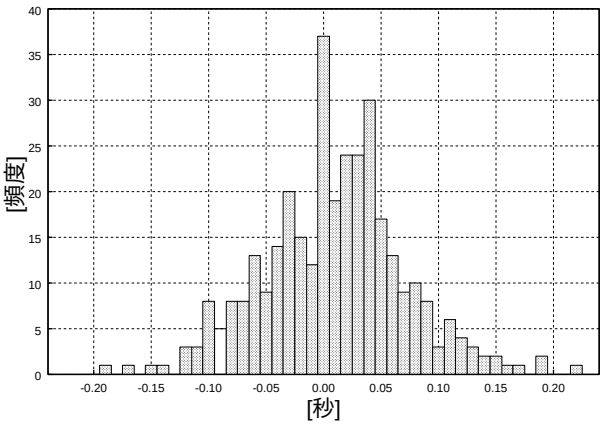

図 3　音節継続時間の差の分布図 (自動ラベル-手動ラベル)

図 3 から，音節開始時間の差の頻度として，最大なものは，-0.05〜0.05 の範囲であることが分かる．

図 3 において，自動ラベルと手動ラベルで大きく異なったデータを以下に示す．

「結婚する」「ke-q」-191ms

「歩いていた」「ta」217ms

5.3　オピニオン評価の実験結果

オピニオン評価の全被験者の平均を表 8 に示す．

表 8　オピニオン評価の実験結果

	オピニオンスコア
自然音声	4.58
手動ラベル	3.67
自動ラベル	3.44

表 8 から，自動ラベルを用いた合成音声のオピニオンスコアは，自動ラベルを用いた合成音声より，やや低い値であるが，あまり差がないことがわかる．そして，高い品質の合成音声が作成できたことが確認できる．しかし，自然性の面で自然音声との差があることが分かる．

5.4　対比較実験の実験結果

5.4.1　手動ラベルと自動ラベルを用いた合成音声との対比較

手動ラベルと自動ラベルを用いた合成音声の対比較実験の結果を表 9 に示す．

表 9　手動ラベルと自動ラベルの対比較の実験結果

	手動ラベル (%)	自動ラベル (%)
文節数 100	59.8	40.2

手動ラベルを用いた合成音声との対比較では，自動ラベルを用いた合成音声の方が良い音声と判定された文は 40.2%となった．これより，自動ラベルを用いた合成音声の自然性が，やや劣化しているが，かなり品質の高い音声が得られたことが示された．

5.4.2　自然音声との対比較

手動ラベルを用いた合成音声と自然音声の対比較実験の結果，および自動ラベルを用いた合成音声と自然音声の対比較実験の結果を表 10 に示す．

表 10　自然音声との対比較実験の結果

	自然音声 (%)	自動ラベル (%)
文節数 100	85.0	15.0
	自然音声 (%)	手動ラベル (%)
文節数 100	80.0	20.0

この結果から，自動ラベルを用いた合成音声，手動ラベルを用いた合成音声ともに，自然性の面で，自然音声との差があるものの，高い品質の合成音声が作成できたことが分かる．

6. 考　　察

6.1　音節境界位置の解析

表 7 から，自動ラベルデータは手動ラベルデータに比べ，音節開始時間，音節終了時間ともに，音節境界位置を 20ms 程度早くラベリングすることが分かる．これより，全体の自動ラベルデータに対し，20ms 程度遅くラベリングすることにより，手動ラベルデータを用いた合成音声に近づいた合成音声が作成できるかもしれない．しかし，図 1，図 2，図 3 から，自動ラベルと手動ラベルの音節境界位置の差の頻度数で，最大となるものは，-0.05〜0.05 の範囲であり，全体の自動ラベルデータに対し，補正を行うべきではないかもしれない．

また，モーラ情報を使って，自動音素ラベリングを行った研究 [13] では，特定話者における音節境界位置の自動ラベルと手動ラベルの差の平均値と標準偏差，および音節継続時間の自動ラベルと手動ラベルの差の平均値と標準偏差を表 11 に示す．

表 11　特定話者における音節境界位置と音節継続時間

	平均値 (ms)	標準偏差 (ms)
音節境界位置	0.50	29.49
音節継続時間	-2.71	42.65

表 7 と表 11 から，特定話者と比較すると，不特定話者の音節境界位置の差，音節継続時間の差はともに，1.5 倍程度大きくなっていることが分かる．これは，自動ラベリングするデータが，不特定話者の自動ラベリングとなったために，精度が悪くなったと考えている．

また，図 1，図 2，図 3 より，他の特異点となったデータをみると，音節開始時間の差においては，自動ラベリングが遅いデータとして，中心の音節が「o」で, 直前の音節は「o-u」であるものが多く存在した．音節終了時間の差においては，自動ラベリングが早いデータとして，中心の音節に促音を含むものが見られ，自動ラベリングが遅いデータとして，中心の音節に「o-u」の連続母音を含むものが存在した．そして，音節継続時間の差においては，自動ラベリングが早いデータとして，中心の音節に撥音，促音を含むものが多く存在した．これより，各音節ごとに，細かく値の補正を行うことで，手動ラベルに近づく合成音声の作成が可能と考えている．

6.2　オピニオン評価の解析

オピニオンスコアでは自然音声は 4.58 であるのに対し，手動ラベルを用いた合成音声では 3.67，自動ラベルを用いた合成音声では 3.44 となり，自然音声に及ばなかった．オピニオンスコアが悪くなった原因としては，接続部での音節ごとの音量の違いであると考えている．

また，手動ラベル自動ラベルを用いた合成音声では，自動ラベルを用いた合成音声のオピニオンスコアが大きく低下した音声があった．低下の原因としては，音節の欠如，または過多にあると考えている．例えば「結婚した (ke-q/koN/shi/ta)」が「けこんした (ke/koN/shi/ta)」と「っ」が抜けた音声に聞こえたり，「大量に (tai/ryou/ni)」が「たいいりょうに (tai/i/ryou/ni)」と「い」が余計に入った音声がある．

6.3　対比較実験の解析

対比較実験では，自然音声との比較では，自然音声の自然性が高いことが分かる．これも，接続部での音量の違いにあると考えている．

また，表 8 より，手動ラベルを用いた合成音声は 59.8%，自動ラベルを用いた合成音声は 40.2%となり，ほとんど差がないことが分かる．これより，自動ラベルを用いた音声でも，かなりの品質が得ることができたと考えている．

6.4　本実験の信頼性

本実験の信頼性を確認するため，音節波形接続型音声合成を文節に適用した研究 [7] で示されたオピニオン評価の実験結果を表 12 に示す．

表 12　先行研究のオピニオン評価の実験結果

	オピニオンスコア
自然音声	4.75
手動ラベル	3.83

表 8 と比較すると，本研究のオピニオンスコアは，先行研究より低いことが分かる．これは，非試験者がが同一でないため低くなったと考えている．

しかし，先行研究における，自然音声との差は 0.17，手動ラベルを用いた合成音声との差は 0.16 となり，ともに同程度，オピニオンスコアが低くなっている．これより，本実験は先行研究と同等の信頼性が得られた実験であると考えている．

6.5　自動ラベルを用いた合成音声の考察

今回の実験では，手動ラベルと自動ラベルで，音節境界位置が大きく異なるデータが少しあった．しかし，オピニオンスコアおよび対比較実験の実験結果から，自動ラベルと手動ラベルを用いた合成音声の品質があまり変わらないことが分かった．これより，自動ラベルを用いた合成音声は，十分実用可能な音声であると考えている．

7. ま　と　め

本研究では，文節発声の音声に対して，自動ラベリングを使用して，合成音声を作成し，どの程度の品質が得られるかを調査した．

音節波形接続方式は，中心の音節・直前の音素・直後の音素・文節中のモーラ位置・文節のモーラ数・文節のアクセント型の一致している音節素片，および連続母音を考慮した音節素片を接続して，音声を合成する．自動音素ラベリングは，HTK を使用し，学習データとして，ATR の単語発話データベース Aset に収録されている，女性話者 10 名の奇数番号データを使用した．

自動ラベルを用いて作成した合成音声は，オピニオンスコアで 3.44 を得られた．対比較実験でも，手動ラベルを用いて作成した合成音声と比較すると，40.2% と差がほとんどなく，精度のよい音声ができたことが分かった．

謝辞　謝辞研究活動に対し，御指導，御教授してくださった鳥取大学工学部知能情報工学科計算機 C 研究室の池原教授と村上助教授に深くお礼申し上げます．加えて，本論文を執筆する

にあたり，参考にさせてい頂いた論文，聴覚実験，そして実験結果の集計，雑用，その他諸等に協力してくださった岡本一輝さん，片山慶一郎さん，松浦祥悟さん，植村和久くんに深く感謝いたします.

文　　献

[1] 石川泰：”音声合成のための韻律制御の基礎”，電子情報通信学会技術研究報告，SP2000-72，pp. 27-34(2000).

[2] 村上仁一，水澤紀子，東田正信：”音節波形接続方式による単語音声合成”，電子情報通信学会論文誌 D-II，Vol. J85-D-II，No. 7，pp. 1157-1165(2002).

[3] 水澤 紀子, 村上 仁一, 東田 正信, ”モーラ数，モーラ位置に基づいた音節波形接続による単語音声合成” 日本音響学会論文集, 2-Q-16, pp.311-312, (1999-10).

[4] 石田隆浩，村上仁一，池原悟：”音節波形接続型音声合成の普通名詞への応用”，電子情報通信学会技術研究報告，SP2002-25，pp. 7-12(2002).

[5] 石田隆浩，村上仁一，池原悟：”モーラ情報とアクセント情報を用いた波形接続型音声合成の普通名詞への応用”，日本音響学会 2003 年春季研究発表会，2-Q-18，pp. 1-409,410(2003)

[6] 加藤琢也，村上仁一，池原悟：”波形接続型音声合成の文節への適用”，日本音響学会 2004 年秋期研究発表会，3-2-12，pp. 1-339，340(2004)

[7] 村上 仁一, 加藤 琢也, 池原 悟, ”音節波形接続型音声合成の文節への適用”, 電子情報通信学会技術研究報告, SP2005-19, pp.43-50, (2005-05).

[8] Nich Campbell and Alan W.Black：”CHATR 自然音声波形接続型任意音声合成システム ”，電子情報通信学会技術研究報告，SP96-7，pp. 45-52Z(1996)

[9] Hidden Markov Model Toolkit(HTK) http://htk.eng.cam.ac.uk/

[10] 中川, 橋本, ”HMM 法とベイズ確率を用いた連続音声の セグメンテーション” 電子情報通信学会論文誌, J72-D-II, pp.1-10 (1989).

[11] 古市, 相澤, 井上, 今井, ”音声認識におけるルールベース 法による話者独立音素セグメンテーション”音響学会誌 55, pp.707-716 (1999).

[12] 鬼山, 荒井, 山下, 北橋, 野村, 溝口, ”知識処理に基づく 音声自動ラベリングシステム”信学技報, SP90-84, pp.53-60 (1991).

[13] 前田，村上，池原，”モーラ情報を用いた音素ラベリング方式の検討”信学技報，SP2001-53(2001-08).