社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　(25)
2003／12／18

モーラ情報を用いた音素ラベリング方式の検討

村上　仁一†　　前田　智広†　　池原　悟†

† 鳥取大学 工学部 知能情報工学科　〒680-8552 鳥取県鳥取市湖山町南 4-101
E-mail: †{murakami,ikehara}@ike.tottori-u.ac.jp, ††99maeta@water.ike.tottori-u.ac.jp

あらまし　音声認識や合成システムなどの音声情報処理の研究において，音素の境界位置を示す音素ラベリングデータは重要である．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いる．しかし現段階の精度はまだ十分ではなく，さらに高い精度が要求されている．ところで音声信号は大きく２つの情報で構成されている．１つはフォルマントで，もう１つはピッチである．この２つの情報を分離するためにケプストラム分析が良く利用される．この分析方法では，低次の項にフォルマントが，高次の項にピッチが抽出される．現在の音声認識や音素ラベリングでは，フォルマント情報が利用されている．つまりケプストラム分析をして低次の項が利用される．しかしケプストラム分析をおこなった場合，ケプストラムの低次の項は，高次の項の影響を受けることが知られている．つまりフォルマントを計算するときに，ピッチが影響することが知られている．一方，ピッチ周波数と単語のモーラ数およびモーラ位置の間に依存関係があることが報告されている．本論文では、このピッチ周波数と単語のモーラ数およびモーラ位置の関係を使うことで，フォルマントにおけるピッチの影響を分離できると仮定した．そして，この関係を使用して自動ラベリングを行えば，音素境界位置の精度は向上すると予想した．この予想を検証するため，始めに母音・促音・撥音を単語のモーラ数および単語のモーラ位置で分類して音素 HMM の学習を行った．次に作成した音素 HMM を使用して単語の音素ラベリングデータを作成した．最後に求められた音素境界位置と，人手によって付与された音素境界位置の差の標準偏差を調べた．この標準偏差を単語のモーラ数およびモーラ位置を使用したときと使用しないときで調べて，本手法の有効性を確認した．

キーワード　ラベリング アクセント モーラ位置 モーラ長 HMM

Segmentation using Mora position and Mora length and Accent

Jin'ichi MURAKAMI†, Tomohiro MAETA†, and Satoru IKEHARA†

† Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University
4-101, Minami Koyamachou, Tottori city, 680-8552 Japan
E-mail: †{murakami,ikehara}@ike.tottori-u.ac.jp, ††99maeta@water.ike.tottori-u.ac.jp

Abstract　In this paper, we consider that the influences between pitch and formant could be separated using the relation of the mora and pitch. And using this relation, the accuracy of the phoneme boundaries will improve. First we trained the phone HMM using mora information. Next we calculate the alignment (phoneme boundaries) using this phone HMM. Last, we evaluated the standard deviation between the obtained phoneme boundaries and human checked phoneme boundaries. As the result of experiments, The effectiveness of mora information is proved.

Key words　alignment, accent, mora length, mora position, HMM

1. はじめに

音声認識や音声合成などの音声情報処理の研究において，音素の境界位置を示す音素ラベリングデータは重要である．現在この音素ラベリングデータは，手作業で作成されており，作成には多大な時間と労力を必要としている．このような作業を軽減するため，大量の音声データを対象に自動的に音素ラベリングを行う自動ラベリングシステムが望まれている．

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いる．HMM 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 [1]，ルールベースを用いる手法 [2]，知識処理に基づく方法 [3] などが過去に報告されている．しかし，現段階の精度はまだ十分ではなく，さらに高い精度が要求されている．

ところで，音声信号は大きく２つの情報で構成されている．１つはフォルマントで舌や喉の動きなどを表し音韻情報を多く含む．もう１つはピッチで声帯の動きを表し個人情報や感情の情報を多く含む．この２つの情報を分離するためにケプストラム分析が良く利用される．この分析方法では，低次の項にフォルマントが，高次の項にピッチが抽出される．

現在の音声認識や音素ラベリングでは，特徴パラメータとしてフォルマント，つまりケプストラムの低次の項が利用される．しかし，ケプストラム分析をおこなった場合，ケプストラムの低次の項は，高次の項の影響を受けることが知られている．つまり，フォルマントを計算するときに，ピッチが影響することが知られている．一方最近の研究において，特定話者の単語発話において，単語のモーラ位置および単語のモーラ数が決まればピッチ周波数がほぼ決まることが知られている [4]．この関係を使用して，単語の音声合成において高い自然性を持った合成音声が得られることが確認されている．

本論文では、このピッチ周波数と単語のモーラ数および単語のモーラ位置の関係を使うことで，フォルマントにおけるピッチの影響を分離できると仮定した．そして，この関係を使用して自動ラベリングを行えば，音素境界位置の精度は向上すると予想した．この予想を検証するため，本研究では，まず，母音・促音・撥音を単語のモーラ数および単語のモーラ位置で分類して音素 HMM の学習を行う．次に，この音素 HMM を使用して単語の音素ラベリングデータを作成する．最後に，求められた音素境界位置と，人手によって付与された音素境界位置の差の標準偏差を調べる．この標準偏差を単語のモーラ数およびモーラ位置を使用したときと使用しないときで調べ，本手法の有効性を確認する．

2. 一般名詞におけるアクセントとモーラ情報とピッチ周波数の関係

本論文では，まず，日本語の一般名詞における単語のアクセントとモーラ位置とモーラ長とピッチ周波数の平均値の関係を調査した．一般名詞には ATR の Aset5240 単語を用いた．アクセントの型は，人の聴覚実験ではなく NHK アクセント辞典 [10] を利用して調査した．ATR の Aset は，日本語の名詞や動詞の単語を含み 5,240 単語で構成されるが，4 モーラの名詞は，1,659 単語あった．

2.1 アクセント型の分布

ATR の Aset の，4 モーラ単語における，アクセント型の単語の頻度を表 1 に示す．この結果から，全単語の 76.1%は 0 型のアクセントであることがわかる．

表 1 4 モーラ単語のアクセントの分布

Table 1 Examples of Accent Distribution

アクセント型	単語数	割合
0 型	1,263 単語	76.1 %
1 型	178 単語	10.7 %
2 型	110 単語	6.6 %
3 型	108 単語	6.5 %
計	1,659 単語	

2.2 アクセント型とモーラ情報とピッチ周波数の関係

一般名詞において，単語のアクセント型とモーラ情報とピッチ周波数の関係を，Xwaves+ [5] を用いて調査した．男性話者 MAU の 4 モーラ単語の 1,659 単語のピッチ周波数の平均と分散を図 1 に示す．この図において，時間軸はモーラ数で正規化したのち計算した．図中の × はピッチ周波数の平均値を，縦線の長さは分散を示している．

また，4 モーラ単語の 0 型の 1,263 単語のピッチ周波数の平均と分散を付録の図 4 に，4 モーラ単語の 1 型の 178 単語のピッチ周波数の平均と分散を付録の図 5 に，4 モーラ単語の 2 型の 110 単語のピッチ周波数の平均と分散を付録の図 6 に，4 モーラ単語の 3 型の 108 単語のピッチ周波数の平均と分散を付録の図 7 に示す．

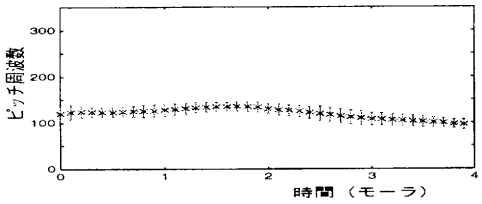

図 1 男性話者 MAU の 4 モーラ単語の 1,659 単語のピッチ周波数の平均と分散

Fig. 1 Pitch frequency for all 4 mora (MAU)

女性話者 FTK の 4 モーラ単語の 1,659 単語のピッチ周波数の平均と分散を図 2 に示す．

また，4 モーラ単語の 0 型の 1,263 単語のピッチ周波数の平均と分散を付録の図 8 に，4 モーラ単語の 1 型の 178 単語のピッチ周波数の平均と分散を付録の図 9 に，4 モーラ単語の 2 型の 110 単語のピッチ周波数の平均と分散を付録の図 10 に，4 モーラ単語の 3 型の 108 単語のピッチ周波数の平均と分散を付録の図 11 に示す．

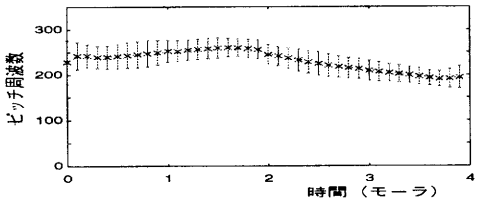

図 2 女性話者 FTK の 4 モーラ単語の 1,659 単語のピッチ周波数の平均と分散

Fig. 2 Pitch frequency for all 4 mora (FTK)

これらの図を見ると，男性話者 MAU では，0 型，1 型，2 型，3 型ともほぼ同様なピッチ周波数の変化をしている（図 4 から図 7）．しかし，女性話者では，0 型，1 型，2 型，3 型それぞれピッチ周波数の変化が異なっている（図 8 から図 11）．しかし全単語の 76.1%が 0 型であるため，4 モーラ単語の平均は，0 型のピッチ周波数とほぼ同様になる (図 2 と図 8)．

以上の結果より，一般名詞において，ピッチ周波数は単語のモーラ数および単語のモーラ位置が決まればほぼ一定であることがわかる．したがって，アクセント型を無視して，モーラ情報とピッチ周波数の依存関係を利用することにより，フォルマントを示すケプストラムの低次の項に対するピッチの影響を分離できると期待できる．

3. モーラ情報を用いた音素ラベリング

アクセントとモーラ情報とピッチ周波数の依存関係 (2. 章) に着目すると，モーラ情報を用いて音素 HMM を作成することで音素ラベルの精度が向上することが期待できる．この仮説を検証するために，始めにモーラ情報を用いて音素 HMM を作成する．次に作成した音素 HMM と Viterbi アルゴリズム [6] を用いて音素境界位置を求める．最後に求められた音素境界位置と人手によって付与された音素境界位置の差の標準偏差を調べる．モーラ情報の有効性の評価は，モーラ情報を使用したときと使用しないときの標準偏差の差から検証する．以下に実験手順を示す．

3.1 実験手順

実験手順を以下に示す．

（1） ラベルファイルの母音・促音・撥音を表している音素記号をモーラ情報を使い分類

（2） データベースを学習データと評価データに分割

（3） 学習データから音素 HMM を作成

（4） 作成された音素 HMM を用いて，評価データの発話内容を既知として，Viterbi decoding を行い，音素境界位置を計算

（5） 計算された音素境界位置と人手によって求められている音素境界位置の差を評価

3.2 ラベルファイルの母音・促音・撥音の分類

データベースの音声ラベルファイルに含まれる母音・促音・撥音を，モーラ情報を使って分類する．具体的には，母音・促音・撥音の前方に単語のモーラ数，後方にモーラ位置情報を付け加えて分類する．分類例を表 2 に示す．

音声ラベルが"akairo" の場合，単語のモーラ数は 4 なので母音の前方に 4 をつけ，後方に各々のモーラ位置をつける．1 番目と 3 番目の音素 a は，分類後は 4a1 と 4a2 という音素に置き換え，モーラ位置が異なるため，異なった音素として扱う．

3.3 評価方法 音素境界位置

評価には，音素境界位置の平均値と標準偏差を使用する．始

表 2 母音と促音の撥音の分類例

Table 2 Examples of Mora length and Mora position

分類前	a	k	a	i	r	o
分類後	4a1	k	4a2	4i3	r	4o4
分類前	a	p	a	a	t	o
分類後	4a1	p	4a2	4a3	t	4o4
分類前	s	e	n	t	a	k	u
分類後	s	4e1	4n2	t	4a3	k	4u4

めに，モーラ情報を使った場合とモーラ情報を使わない場合で音素ラベリングを行う．次に，計算によって求められる音素境界位置と人手によって求められた音素境界位置を比較し，平均値と標準偏差を求める．

図 3 に，発話内容が"taido" である評価データの人手によって求められる音素境界位置 (a_1～a_6) と計算によって求められる音素境界位置 (b_1～b_6) を示す．横軸は時間，縦軸は波形の振幅を表す．

音素境界位置の平均値 E_p の計算式を以下に示す．

$E_p = \frac{\Sigma_i (a_i - b_i)}{n} (i = 1, 2, .., n)$

音素境界位置の標準偏差 σ_p の計算式を以下に示す．

$\sigma_p = \sqrt{\frac{\Sigma_i (a_i - b_i - E_p)^2}{n}} (i = 1, 2, .., n)$

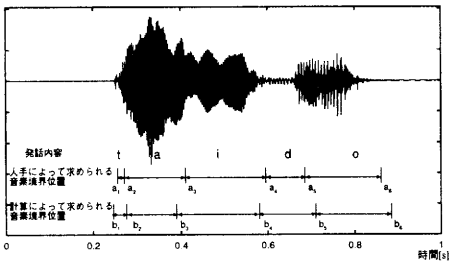

図 3 波形データと音素境界位置

Fig. 3 wave and segmentation data

3.4 実験条件

データベースには ATR の単語発話データベース Aset の 5,240 単語を使用し，奇数番を学習データに，偶数番を評価データとする．学習データ，評価データともに音素数は約 15,500 で母音数は約 8,000 である．使用する音声データは全て，人手によって音素境界位置が付与されている．

評価実験は，男性話者 10 名と女性話者 10 名で行う．ケプストラムの計算や音素 HMM の学習や自動ラベリングなどには HTK [6] を使用する．音響モデルにはラベリングの精度を高くするために Full-covariance HMM を使用して実験を行う．その他の実験条件を表 3 に示す．

モーラ情報を使って母音・促音・撥音を分類すると，音素の種類は，26 種類から約 160 種類に増加する．しかし，学習データが不十分であるために学習ができない音素 HMM がある．そのため，評価実験で使用される音素 HMM は約 80 種類となった．また，学習データが不十分で音素 HMM が作成できない音素を含む評価データは，評価から除外した．

4. 実験結果

4.1 音素境界位置

音響モデルに Full-covariance HMM を用いた場合に，計算した音素境界位置と人手によって付与された音素境界位置の差の平均値と標準偏差を表 4，5 に示す．表 4 は男性話者 10 名の結果で，表 5 は女性話者 10 名の結果である．

表 3 実験条件

Table 3 Experimental Conditions

標本周波数	16kHz	学習 DB	2,620 単語
分析窓	Hamming 窓	音素数	約 15,500
分析窓長	20ms	母音数	約 8,000
フレーム周期	5ms	評価 DB	2,620 単語
音響モデル	4 状態 3 ループ	音素数	約 15,500
mixture	3	母音数	約 8,000
特徴ベクトル	16 次 MFCC+		
	対数パワー (計 17 次)		

表 4 男性話者の音素境界位置 (Full)

Table 4 Phone Boundary Positions - Male - (Full)

DB	モーラ無し			モーラ有り		
	調査音素数	平均値	標準偏差	調査音素数	平均値	標準偏差
	n	$E_p(ms)$	$\sigma_p(ms)$	n	$E_p(ms)$	$\sigma_p(ms)$
MAU	18144	-1.62	22.12	16875	-2.35	21.60
MHT	18141	0.31	24.89	16663	0.18	22.66
MMS	18150	-2.35	20.75	16768	-1.98	19.39
MMY	18035	-1.70	21.91	16833	-1.72	21.29
MNM	18148	-1.37	21.84	16913	-0.89	20.67
MSH	18143	-3.02	24.62	16828	-2.57	23.55
MTK	18157	-1.30	24.68	16466	-1.80	24.05
MTM	18159	-1.24	23.16	16760	-2.86	20.95
MTT	18165	-2.42	21.78	16887	-2.28	19.30
MXM	18153	-1.63	22.14	16770	-2.02	20.96
平均		-1.63	22.79		-1.83	21.44

表 5 女性話者の音素境界位置 (Full)

Table 5 Phone Boundary Positions - Female - (Full)

DB	モーラ無し			モーラ有り		
	調査音素数	平均値	標準偏差	調査音素数	平均値	標準偏差
	n	$E_p(ms)$	$\sigma_p(ms)$	n	$E_p(ms)$	$\sigma_p(ms)$
FAF	18162	-0.91	28.62	16992	-0.84	23.07
FFS	18090	-0.79	30.60	16823	-0.17	24.32
FKM	18164	-0.34	26.16	17059	-1.23	21.50
FKN	18143	1.67	29.51	16990	1.81	28.62
FKS	18145	-1.74	25.73	16946	-1.34	22.71
FMS	18158	-2.91	24.65	17033	-1.22	21.69
FSU	18040	-0.32	26.86	16929	-0.37	23.72
FTK	18157	-2.20	24.25	16926	-1.73	21.79
FYM	18129	0.91	28.02	16973	1.66	24.35
FYN	18148	-1.01	32.09	17146	-0.22	30.10
平均		-0.76	27.65		-0.37	24.19

表 4～5 より，全ての話者においてモーラ情報を使用することで，標準偏差が小さくなっていることがわかる．

Full-covariance HMM を用いた場合，モーラ情報を使用することにより，男性話者の場合の標準偏差は平均 1.4ms(22.79 – 21.44) 精度が向上した．また女性話者の場合の標準偏差は約 3.5ms(27.65 – 24.19) 精度が向上した．

なお，Full-covariance HMM は，Diagonal-covariance HMM と比較するとモーラ情報を使用することにより，男性話者の場合の標準偏差は平均約 1.7ms(24.43 – 22.72) 精度が向上した．また女性話者の場合の標準偏差は約 4.4ms(29.49 – 25.09)

精度が向上した．FTK を除き，音響モデルに Full-covariance HMM を使用した方が，Diagonal-covariance HMM より精度は良かった．

以上の実験により，モーラ情報の有効性が確かめられた．

5. 考　　察

5.1 音韻間の解析

各話者における音韻毎の音素境界位置の精度を調べた．MAU の場合を表 6，FTK の場合を表 7 に示す．なお，いずれも音響モデルは Full-covariance HMM を使用した．実験条件は表 3 と同等である．

これらの表 6 ～ 7 をみると，特に母音と母音の音素境界位置の精度が良くなっていることがわかる．母音と母音の音素境界位置の標準偏差は，男性話者では約 14.8ms (54.52 − 40.05) 向上し，女性話者では約 27.8ms (70.98 − 43.73) 精度が向上した．しかし，他の音素境界においては精度の向上はあまり見られない．

この理由として，母音にはピッチが含まれるため，モーラ情報を利用することにより，ピッチの影響が分離できるのに対し，子音では，ピッチがないため，モーラ情報を使用しても効果がなかったと考えてられる．

表 6　MAU の音韻間の音素境界位置 (Full)

Table 6　Phone Boundary Positions - MAU - (Full)

音素間	モーラ無し		モーラ有り	
	調査音素数 n	標準偏差 $\sigma_p(ms)$	調査音素数 n	標準偏差 $\sigma_p(ms)$
母音-母音	1351	54.03	1266	40.87
母音-半母音	846	11.51	735	13.52
母音-鼻音	442	9.88	402	10.05
母音-無音	2470	16.36	2322	21.01
母音-摩擦音	579	15.42	513	16.80
母音-促音	565	17.00	554	27.76
母音-有声破裂音	462	12.26	419	13.45
母音-無声破裂音	1175	12.73	1084	11.29
半母音-母音	1400	15.40	1271	15.55
鼻音-母音	756	15.32	700	18.09
無音-母音	339	8.15	303	7.87
摩擦音-母音	1265	17.20	1171	18.12
促音-母音	12	37.94	12	20.61
有声破裂音-母音	700	14.88	650	17.81
無声破裂音-母音	2036	12.10	1890	12.94

5.2 全音素にモーラ情報を適用

表 4 ～ 5 の実験では，母音・促音・撥音を単語のモーラ数および単語のモーラ位置で分類して音素 HMM の学習を行った．この節では全ての音素をモーラ情報を使って分類した場合の音素境界位置の精度を調べた．データベースは MAU, MMY, FAF, FTK を使用し，音響モデルは Diagonal-covariance HMM を使用した．その他の実験条件は表 3 と同様である．

実験結果を表 8 に示す．いずれの話者も，モーラ情報を使って母音・促音・撥音を分類した場合より，音素境界位置の精度の向上は見られず，4 名の平均では標準偏差が 0.47ms(23.05 − 22.58) 増加した．

この原因として，モーラ情報を使って子音を分類して音素ラベリングを行っても，子音にはピッチが含まれないために，音素境界位置の精度は向上しなかったと考えられる．そして 1 個あたりの音素 HMM の学習データが減少し，そのため HMM の精度が低下し，音素境界位置の精度が低下したと考えられる．

5.3 半連続分布型 HMM を用いたラベリング精度の検討

表 4 ～ 5 より，全ての実験において，単語のモーラ数と単語のモーラ位置を用いることにより，音素境界位置および音素継続時間の精度向上が認められた．しかし，単語のモーラ情報を用いて HMM を作成した場合と，用いずに HMM を作成したときでは，HMM のパラメータの数が異なるため，公平な比較にはならない．そこで，全ガウス分布の数を固定できる半連続分布 HMM [11] を用いて同様な実験を行った．

なお，HTK を用いて半連続分布 HMM を作成する際，連結学習が必須になる．しかし，連結学習をおこなうと自動ラベリングの精度が低下することが一般的に知られている．そのため，HMM の学習は，半連続型 HMM において連結学習を行った後で，再び音素ごとに Baum-Welch 学習を行った．結果を表 9 にまとめる．なお，半連続型 HMM のガウス分布の数は 256 とし，Diagonal-covariance HMM で計算した．

この結果をみると，女性話者ではモーラ情報をもちいることにより，セグメンテーションの精度が向上していることがわかる．しかし，男性話者では精度が低下した．

現在の HTK では半連続分布型 HMM を作成するときに連結学習が必要になる．これが問題になったと考えられる．今後，連結学習を行わずに半連続型 HMM を学習する方法を検討する必要がある．

表 7　FTK の音韻間の音素境界位置 (Full)

Table 7　Phone Boundary Positions - FTK - (Full)

音素間	モーラ無し		モーラ有り	
	調査音素数 n	標準偏差 $\sigma_p(ms)$	調査音素数 n	標準偏差 $\sigma_p(ms)$
母音-母音	1350	60.76	1272	40.77
母音-半母音	846	13.77	744	15.23
母音-鼻音	442	7.99	399	9.57
母音-無音	2470	15.68	2326	18.52
母音-摩擦音	579	23.91	516	23.89
母音-促音	565	16.59	554	20.43
母音-有声破裂音	462	11.43	417	12.04
母音-無声破裂音	1174	17.39	1083	16.27
半母音-母音	1415	15.16	1295	17.06
鼻音-母音	756	18.79	698	17.47
無音-母音	339	8.16	302	7.88
摩擦音-母音	1256	15.73	1165	16.27
促音-母音	12	49.47	12	28.24
有声破裂音-母音	699	12.14	647	13.19
無声破裂音-母音	2037	14.60	1895	15.72

表 8　全音素にモーラ情報を適用したときの音素境界位置 (Diagonal)

Table 8　Phone Boundary Positions (Diagonal)

DB	全音素分類			母音・促音・撥音分類		
	調査音素数 n	平均値 $E_p(ms)$	標準偏差 $\sigma_p(ms)$	調査音素数 n	平均値 $E_p(ms)$	標準偏差 $\sigma_p(ms)$
MAU	15126	-2.51	23.08	17771	-1.47	22.56
MMY	15048	-1.44	22.45	17663	-0.86	22.12
FAF	15109	-1.47	23.97	17781	-0.30	23.43
FTK	15147	-1.07	22.69	17802	-0.67	22.22
平均		-1.62	23.05		-0.83	22.58

5.4 アクセント情報を用いたラベリング精度の検討

モーラ情報にアクセント位置の情報を加えた場合のセグメンテーションの精度を調査した．アクセント位置は，NHK アクセント辞典 [10] を利用した．また，音素数が多くなるため，全ての HMM のガウス分布が共通な半連続型 HMM [11] を使用した．他の実験条件は 5.3 と同一である．結果を表 10 に示す．

表 10 と表 9 を比較すると，アクセント情報を利用することにより，精度が向上していることがわかる．特に女性話者において精度が向上している．これは，女性話者のほうがアクセン

表 9　半連続型 HMM におけるセグメンテーションの結果 (Diagonal)

Table 9　Result of Semi-Continuous HMM (Diagonal)

DB	モーラ無し			モーラ有り		
	調査音素数	平均値	標準偏差	調査音素数	平均値	標準偏差
	n	$E_d(ms)$	$\sigma_d(ms)$	n	$E_d(ms)$	$\sigma_d(ms)$
MAU	17334	-2.53	20.61	15959	-2.37	21.86
MMY	17700	-2.91	19.83	15509	-2.31	19.74
MTK	17347	-0.40	20.68	15534	-1.04	22.43
男性平均		-1.95	20.37		-1.91	21.34
FTK	17395	-1.78	22.24	16218	-2.01	21.33
FYN	17598	-1.28	28.70	16372	0.13	28.45
FAF	17700	-2.53	26.73	17127	-2.92	24.53
女性平均		-1.86	25.89		-1.60	24.77

表 10　アクセント情報を用いたセグメンテーションの結果 (Diagonal)

Table 10　Result of Accent (Diagonal)

DB	調査音素数	平均値	標準偏差
	n	$E_d(ms)$	$\sigma_d(ms)$
MAU	12039	-3.30	21.73
MMY	11825	-2.15	19.32
MTK	9370	-0.89	20.11
男性平均		-2.11	20.39
FTK	14063	-2.22	19.93
FYN	15249	-2.46	27.40
FAF	15260	-2.84	24.13
女性平均		-2.51	23.82

ト型の違いにおいてピッチ周波数の変化の様子が大きく異なるためと考えられる．

5.5　男性話者と女性話者の比較

女性話者は男性話者に比べて，アクセントやモーラ情報を使うことによる音素境界位置の認識精度は向上した（表 4 ～ 5 参照）．このことから，アクセントやモーラ情報は，男性話者よりも女性話者に効果があることがわかった．この原因について以下の理由が考えられる．

通常，女性話者のピッチ周波数は，男性話者のピッチ周波数に比べて高い．そのため，ケプストラムの低次の項（フォルマント）がピッチに大きく影響される．この影響がモーラ情報を使用することで分離できた．その結果，女性話者の方が男性話者より効果が大きかったと考えられる．

5.6　モーラ位置およびモーラ長およびアクセントの有効性

表 4 ～ 5 より，単語のモーラ数と単語のモーラ位置を用いることにより，音素境界位置の精度向上が認められた．特に母音と母音の間の境界位置の精度が向上した．したがってピッチ周波数と単語のモーラ数および単語のモーラ位置の依存関係を使うことで，フォルマントにおけるピッチの影響が分離し，その結果，音素境界位置の精度は向上したと考えられる．

しかし，全音素のガウス分布の数を共通にした半連続型 HMM を使用した実験では，女性話者において有効性が示せたが，男性話者では有効性が示せなかった．この原因として連結学習による影響が考えられる．また，モーラ情報にアクセントを加えて HMM のモデルを作成することにより，さらに有効性が見られた．今後，さらに triphone モデルなどでも精度の調査を行っていきたい

5.7　ラベリングの精度

音響モデルに Full-covariance HMM を使用した場合，音素境界位置のずれから求めた標準偏差は，男性話者で 21.44ms，女性話者で 24.19ms であった．人手によって求められる音素境界位置のゆらぎは 5ms 程度といわれているので，さらに精度を向上させる必要があろう．

しかし，自動ラベリングの応用として，音素片を繋いで合成する音声合成がある．[7] [8]．この方法において，手動ラベリングを利用して合成した音声と自動ラベリングを利用して合成した音声を比較した場合，音質に差が殆んどないことが示されている [9]．したがって，音声合成における自動ラベリングの精度は，本論文で示した値で実用上問題がないと考えられる．

6.　ま と め

本論文では，単語のモーラ数および単語の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まることを利用して，母音・促音・撥音の HMM を，単語のモーラ数および単語のモーラ位置で分類して学習を行い，単語の音素境界位置を求めた．そして，モーラ情報を使用した場合と使用しない場合で，人手によって求められている音素境界位置と計算によって求められた音素境界位置を比較し標準偏差を求めた．その結果，音響モデルに Full-covariance HMM を使用した場合，男性話者 10 人の標準偏差は約 1.4ms 向上し，女性話者 10 人の標準偏差は，約 3.5ms 精度が向上した．したがってモーラ情報を利用することによる有効性が得られた．

各音素境界位置ごとにモーラ情報の有効性を調べたところ，もっとも有効であったのは，母音と母音の音素境界位置であった．また，モーラ位置の情報にアクセントを加えて自動ラベリングをおこなったところ，さらに精度が向上することが示された．そして，男性話者と女性話者の実験結果を比較したところ，モーラ情報は男性話者に比べ女性話者の方が有効であることがわかった．

今後，最適な実験条件のパラメータの検討や triphone モデルやモーラ情報が不特定話者の音素ラベリングに有効であるか調べる必要がある．

謝辞

現在鳥取大学大学院修士課程 2 年の石田隆浩君に，ATR, Aset の平均ピッチ周波数を計算してもらいました．感謝いたします．

文　献

[1] 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" 信学論, J72-D-II,1-10(1989)

[2] 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)

[3] 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

[4] 水澤, 村上, 東田, " 音節波形接続による単語音声合成" 信学技報, SP99-2(1999-05)

[5] Introducing ESPS/waves+ with EnSigTM Entropic Research Laboratory,Inc.

[6] HTK Ver2.2 referene manual, 1997 Cambridge University

[7] N.Campbell and A.Black, "CHATR:自然音声波形接続型任意音声合成システム," 信学技報, SP96-7, pp45-52(1996.5)

[8] 村上, 水澤, 東田, " 音節波形接続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 (2002-07)

[9] 石田, 村上, 池原" 音節波形接続型音声合成の普通名詞への応用", 信学技報, SP2002-25 pp.7-12 (2002-05)

[10] NHK 放送文化研究所, "NHK 日本語発音アクセント辞典 ", 日本放送出版協会ISBN:4140111127 (1998-04)

[11] X.D.Huang, Y.Ariki, M.A.Jack, "HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION", Edinburgh University Press, (1990)

付録

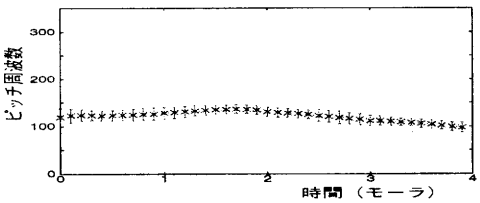

図 4　男性話者 MAU の 4 モーラ 0 型の 1,263 単語のピッチ周波数の平均と分散
Fig. 4　Pitch frequency for 4 mora type 0 (MAU)

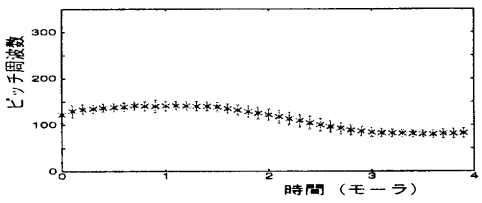

図 5　男性話者 MAU の 4 モーラ 1 型の 178 単語のピッチ周波数の平均と分散
Fig. 5　Pitch frequency for 4 mora type 1 (MAU)

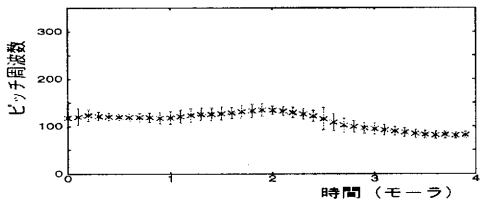

図 6　男性話者 MAU の 4 モーラ 2 型の 110 単語のピッチ周波数の平均と分散
Fig. 6　Pitch frequency for 4 mora type 2 (MAU)

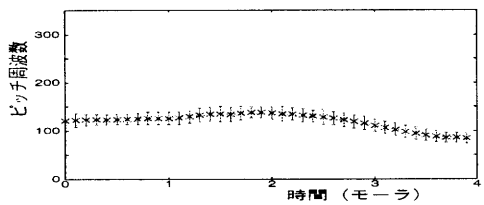

図 7　男性話者 MAU の 4 モーラ 3 型の 108 単語のピッチ周波数の平均と分散
Fig. 7　Pitch frequency for 4 mora type 3 (MAU)

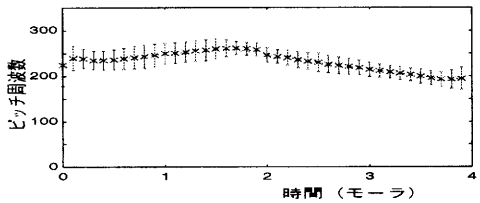

図 8　女性話者 FTK の 4 モーラ 0 型の 1,263 単語のピッチ周波数の平均と分散
Fig. 8　Pitch frequency for 4 mora type 0 (FTK)

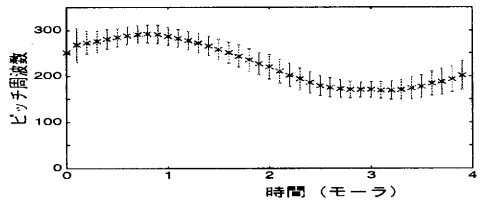

図 9　女性話者 FTK の 4 モーラ 1 型の 178 単語 のピッチ周波数の平均と分散
Fig. 9　Pitch frequency for 4 mora type 1 (FTK)

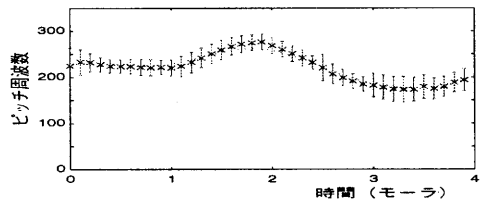

図 10　女性話者 FTK の 4 モーラ 2 型の 110 単語のピッチ周波数の平均と分散
Fig. 10　Pitch frequency for 4 mora type 2 (FTK)

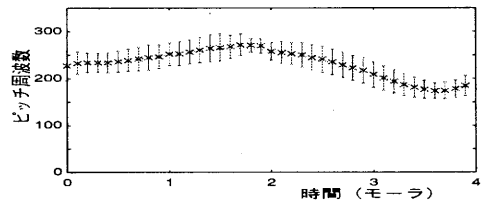

図 11　女性話者 FTK の 4 モーラ 3 型の 108 単語のピッチ周波数の平均と分散
Fig. 11　Pitch frequency for 4 mora type 3 (FTK)