音声言語情報処理 39－18
(2001. 12. 20)

アクセント句境界情報を利用したN-gram言語モデルの高精度化

寺尾 真† 峯松 信明†† 広瀬 啓吉†††

† 東京大学 大学院 工学系研究科
†† 東京大学 大学院 情報理工学系研究科
††† 東京大学 大学院 新領域創成科学研究科
〒113-0033 東京都文京区本郷 7-3-1

E-mail: {terao,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし 現在の大語彙連続音声認識システムでは、音声の音韻的特徴が主に用いられ、韻律的特徴はほとんど利用されていない。そこで本研究では、韻律的特徴を利用してN-gram言語モデルを高精度化する手法を提案する。本手法では、韻律的特徴の1つであるアクセント句境界の有無に応じた2種類のN-gram言語モデルを構築し、これを句境界の有無によって使い分ける。しかし、この2種類の言語モデルを直接単語レベルで構築することは、音声データの量が少ないため非現実的である。そこで、句境界の有無による品詞遷移の特徴の違いを利用して、これらをベースとなる通常の言語モデルから構築する手法を考案した。提案する言語モデルを、ベースとなった言語モデルと比較評価したところ、正解の句境界によって提案言語モデルの学習および評価を行ったときに、約11%のパープレキシティの低下を実現した。また、音声データから自動抽出した句境界を用いても約8%改善された。さらに、提案言語モデルの学習と評価との間で話者が異なる場合でも、約6%の性能向上が得られた。

キーワード 言語モデル、韻律、アクセント句境界、品詞遷移、連続音声認識

Improvement of N-gram Language Models Using Accent Phrase Boundaries

Makoto TERAO†, Nobuaki MINEMATSU††, and Keikichi HIROSE†††

† Graduate School of Engineering, University of Tokyo
†† Graduate School of Information Science and Technology, University of Tokyo
††† Graduate School of Frontier Sciences, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

E-mail: {terao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract Current continuous speech recognition systems make much use of segmental features but little use of prosodic features. This paper proposes a novel method to integrate prosodic boundary information into N-gram-based language modeling. In this method, two types of language sub-models are built. One characterizes word transitions crossing accent phrase boundaries and the other not crossing the boundaries. To realize these two sub-models directly from a speech corpus, its size should be comparable to a text corpus used for N-gram model training. However, the preparation of such a large speech corpus is not realistic. To solve this problem, we focus upon transition of words in terms of their part-of-speech (POS), and differences in POS transition crossing and not crossing the boundaries are used to generate the two sub-models. Through experiments, the proposed model showed 11% perplexity reduction given the correct boundary position, and 8% reduction with the automatically extracted boundaries. Even when test speech samples were spoken by another speaker than the speaker used in characterizing the POS transitions, 6% reduction was observed.

Key words language model, prosody, accent phrase boundary, transition of part-of-speech, continuous speech recognition

1. はじめに

現在の大語彙連続音声認識システムでは、音声の音韻的特徴が主に利用され、韻律的特徴はむしろ積極的に排除される傾向にあった。しかし、韻律情報は文字言語には無い音声に固有の特徴であり、人間の音声知覚過程において重要な役割を担っているであろうことは明らかである。従って、計算機でより高度な音声認識を行うためには韻律情報の有効な利用は欠かせないと考えられる。しかし、これまでの韻律を利用する研究は、孤立単語音声の認識 [1] やキーワードスポッティング [2] などのように、タスクが限定されているものがほとんどである。

一方、大語彙連続音声認識に韻律を利用した研究としては、[3] が挙げられる。[3] では韻律情報を用いて、探索のビーム幅の動的な制御、および音響モデルの選択を行う手法を提案している。その結果、ビーム幅の動的制御によって計算時間と消費メモリの大幅な削減が達成され、音響モデルの選択によって認識率が向上したと報告されている。このことから、連続音声認識においても韻律情報を適切に用いれば効果があると考えられる。

本稿では、韻律的特徴を連続音声認識に利用する方法として言語モデルに焦点を当て、アクセント句境界を利用した N-gram 言語モデルの高精度化手法を提案し、提案言語モデルのパープレキシティによる評価実験について報告する。

2. アクセント句境界情報を利用した言語モデル

2.1 概　　要

本研究では、連続音声認識の性能向上を目的として、アクセント句境界情報を利用した N-gram 言語モデルの高精度化手法を提案する。これを連続音声認識に組み込む場合は、従来の認識の枠組みに加えて、図 1 のように入力音声を韻律分析してアクセント句境界を抽出し、これによって言語モデルを制御する部分が加わる。ここで、アクセント句とは 1 つのアクセント核を持つ単語のひとまとまりのことで、基本周波数の山にほぼ相当する韻律情報である。

提案手法では、アクセント句境界の有無に応じた 2 種類の N-gram 言語モデルを構築し、これを句境界の有無によって使い分ける（図 2）。この手法の狙いは、アクセント句境界の有無によって日本語の言語的な遷移の性質が異なると仮定し、これを利用することにある。なお、本稿では言語モデルとしてバイグラムを扱っている。

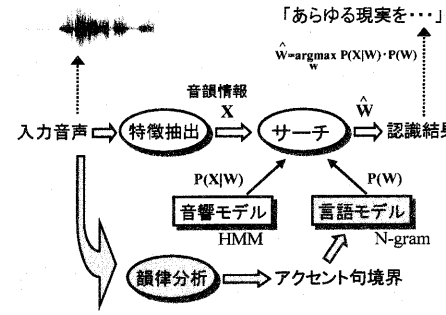

図 1　連続音声認識に韻律情報を利用する枠組み

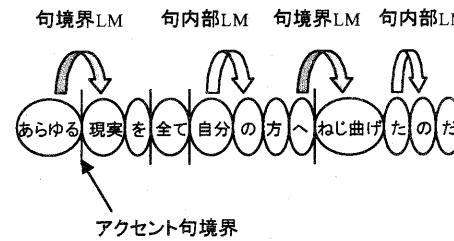

図 2　提案する言語モデル

2.2 提案言語モデル構築の問題点と解決方法

テキストベースで構築できる通常の言語モデルと異なり、図 2 のような 2 種類の言語モデルを構築するためには、句境界情報が得られる学習データが必要となる。したがって、このような言語モデルの学習は書き起こしテキスト付きの音声データベースから行うことになる。しかし、現在利用可能な音声データの量は、新聞記事データベースなどが整備されているテキストデータに比べて非常に少ないため、句境界有無別の言語モデルを単語レベルで直接構築することは非現実的である。

そこで、句境界の有無による言語的な違いは品詞遷移の特徴という形で学習する。そして、この品詞遷移の特徴の違いを利用して、ベースとなる通常の言語モデルから 2 種類の言語モデルを構築するという方法をとる。この考え方を示したのが図 3 である。

2.3 句内部と句境界における品詞遷移特徴の調査

図 3 の構築方法は、アクセント句境界の有無に応じた言語的な遷移性質の違いが、品詞の遷移という形で現れていることを前提としている。そこで句内部、及び句境界における品詞の遷移確率を調査した。調査するテキストデータとしては、ATR の 503 文 [4] を利用した。アクセント句境界の位置はデータベー

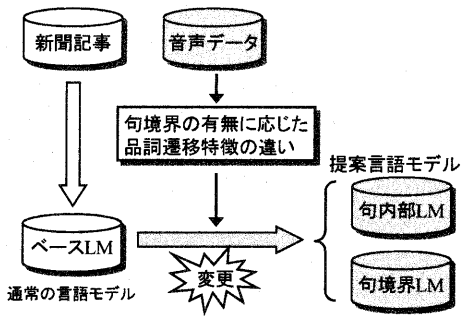

図 3　提案言語モデル構築の概念図

スで提供されている話者 MYI の正解ラベルを使用した。また、品詞体型としてはテキストを茶筌 version2.02 で形態素解析した結果を用いた。この調査結果の中から、「名詞」「動詞」「助詞」「副詞」の 4 品詞間の句内部及び句境界での遷移確率を示したのが、表 1、表 2 である。これらより、アクセント句境界の有無によって品詞遷移の様子に明確な違いが認められる。したがって図 3 のように、品詞遷移の特徴の違いをうまく利用することで、通常の言語モデルから句境界有無別の言語モデルを構築できると考えられる。

表 1　句内部での品詞の遷移確率 (%)

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	8.9	5.2	67.5	0.1
	動詞	6.2	12.7	43.8	0.0
	助詞	6.8	47.9	36.9	0.3
	副詞	2.5	15.0	60.0	0.0

表 2　句境界での品詞の遷移確率 (%)

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	71.1	13.4	1.4	2.8
	動詞	85.6	5.7	1.1	4.0
	助詞	51.1	34.3	0.2	6.1
	副詞	59.6	28.8	0.0	1.4

2.4　句境界の有無に応じた言語モデルの構築方法

以下では、ベースとなる言語モデルからアクセント句境界の有無に応じた 2 種類の言語モデルを構築する方法を具体的に述べる。

（1）　句境界の有無別に品詞遷移をカウント

まず、句境界情報が得られる音声データを用いて、データ中の全ての単語遷移における品詞遷移がどのようになっているのかを句境界の有無別に数え上げる。

この結果、例えば表 3 のような句境界有無別の品詞遷移のカウント表ができる。この例は、代名詞から格助詞への単語遷移が学習データ中に句内部で 90 箇所、句境界では 10 箇所存在したことなどを示している（ただし、表 3 は実際のデータではない）。

なお、本稿では品詞体系として茶筌 version2.02 の解析結果をさらに表 4 の 26 種類に区分してこのようなカウントを行った。

表 3　句境界有無別の品詞遷移のカウント例

品詞遷移	句内部	句境界
代名詞 → 格助詞	90	10
·····	··	··
格助詞 → 名詞一般	10	40
·····	··	··

表 4　26 種類の品詞区分

名詞・一般	動詞・非自立	助詞・並立助詞
名詞・代名詞	動詞・その他	助詞・連体化
名詞・副詞可能	形容詞	助詞・その他
名詞・サ変接続	副詞・一般	助動詞
名詞・形動語幹	副詞・助詞接続	記号
名詞・非自立	連体詞	読点
名詞・接尾	助詞・格助詞	句点
名詞・その他	助詞・接続助詞	その他
動詞・自立	助詞・係助詞	

（2）　ベース言語モデルのバイグラムカウントを分配

バイグラムカウントとは、言語モデルの学習テキスト中に現れる連続した単語の 2 つ組みの出現回数を全て数え上げたものである。例えば、「私、は」という 2 つ組みが学習テキスト中に 1000 回出現した、などとなる。バイグラム言語モデルはこのようなバイグラムカウントから構築される。

そこで提案手法では、ベースとなる通常の言語モデルのバイグラムカウントの値を、句内部用バイグラムカウントと句境界用バイグラムカウントとに分配し、これらを元に句境界有無別の言語モデルをそれぞれ構築する。ある 2 つ組みのバイグラムカウントの分配は、前節でカウントした句内部と句境界のそれぞれにおける対応する品詞遷移の個数に比例して行う。

例えば、ベースとなるバイグラムにおいて「私、は」のバイグラムカウントが 1000 であったとする。この 2 つ組みは表 4 の品詞区分では、「代名詞、格助詞」という遷移である。ここで学習用音声データ中の、品詞遷移のカウント結果が表 3 であったとすると、「代名詞、格助詞」という遷移は句内部で 90 箇所、句境界で 10 箇所あることがわかる。バイグラムカウントを

これに比例して分配するので、句内部用の「私、は」のバイグラムカウントは $1000 \times 90/(90+10) = 900$、句境界では $1000 \times 10/(90+10) = 100$、となる。

以上の作業は、ベースとなるバイグラムカウントのうちいくつが句内部から発生し、いくつが句境界から発生したのかを品詞遷移の特徴から推定していることになる。この方法では、例えば同じ「代名詞、格助詞」という品詞の遷移であれば、それが実際にどのような単語であっても句内部および句境界から同じ比率で発生する、という近似が行われている。

以上で述べた方法によってベースのバイグラムカウントを分配し、句境界の有無に応じた2種類の言語モデルを構築する流れを示したのが図4である。

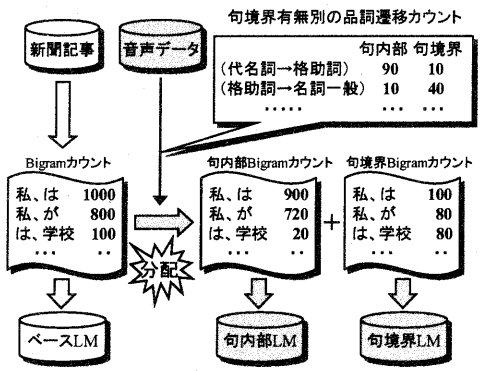

図4 バイグラムカウントの分配による2種類の言語モデルの構築方法

3. パープレキシティによる評価実験

パープレキシティ（以下PP）によって、提案言語モデルをベースとなった言語モデルと比較評価した。提案言語モデルによるPPの計算は、句境界の有無に応じて2種類の言語モデルを切り替えて行った。品詞遷移特徴の学習元となるデータ、及びPPの計算に用いる評価データとしてはATRの503文[4]を用いた。

3.1 ベースとなるバイグラム

ベースとなるバイグラム言語モデルは、97年度毎日新聞記事テキストから構築した。語彙数は20K、形態素解析には茶筌version2.02を使用し、バイグラムの構築にはGood Turing discountingを用いた。

3.2 正解句境界を用いた場合

まず、アクセント句境界情報としてデータベースで提供されている話者MYIの正解ラベルを利用して、提案言語モデルの学習及び評価を行った。

表5は、503文全文で品詞遷移の特徴を学習し、503文全文でPPの計算を行った、品詞遷移学習データについてクローズな実験結果である。これより、提案した言語モデルはベースとなった言語モデルに対して、PPが約11%低下しており、有効であることが分かる。また、表中の「句内部」及び「句境界」の結果は、句内部の遷移のみに対するPP、及び句境界の遷移のみに対するPPをそれぞれ計算した結果である。提案言語モデルは句内部では約9%、句境界では約15%の効果があることが分かる。従って、句内部においても句境界においても提案言語モデルは効果をあげている。

また表6は、503文中の453文で品詞遷移の特徴を学習し、残り50文でPPの計算を行った、品詞遷移学習データについてオープンな実験結果である。実験は、学習データと評価データの組合わせを変えて10通り行いその平均を求めた（Cross Validation）。品詞遷移学習データについてオープンになることで効果は落ちているが、PPは約9%改善されている。

表5 MYI正解句境界503文で学習及び評価

	全文	句内部	句境界
ベース言語モデル	117.0	25.56	2664
提案言語モデル	104.1	23.32	2253
改善率	11.0%	8.76%	15.4%
バイグラムヒット率	95.35%	97.65%	90.60%

表6 MYI正解句境界453文で学習、MYI正解句境界50文で評価

	全文	句内部	句境界
ベース言語モデル	117.4	25.66	2752
提案言語モデル	107.1	23.93	2408
改善率	8.77%	6.74%	12.5%

3.3 句境界を自動抽出した場合

提案言語モデルを実際に連続音声認識システムで利用するためには、音声データからアクセント句境界を自動抽出しなければならない。そこで、音声から自動抽出された句境界に対しても提案言語モデルが有効であるかどうかを前節同様にPPで評価した。

3.3.1 アクセント句境界の自動抽出方法

まず、対象とする音声から[5]によってアクセント句境界の時間を抽出した。[5]は、モーラを単位としたモーラ遷移確率モデルによってアクセント型をHMMでモデル化して、アクセント句境界を検出する手法である。次に、この音声に対して強制切り出しを行い、各形態素境界の時間を求めた。ここで、形態素境界の前後それぞれ40msの範囲に、[5]によって自動抽出されたアクセント句境界が存在するときに、この形態素間にはアクセント句境界が存在すると判定

した。なお1モーラを約100msとしたときにその半分弱の時間長として40msという時間幅を設定した。

また、[5]によるアクセント句境界の検出精度は、話者MYIに対して表7であった。表7は、正解の句境界数に対する検出された句境界の数、及び挿入誤りの数の割合である。

表7　[5]による句境界の自動抽出精度

正解とする時間幅	検出率	挿入誤り率
± 40ms	57%	24%
± 100ms	65%	16%

3.3.2　話者クローズな場合

表8、表9は話者MYIの音声から自動抽出された句境界に基づいて提案言語モデルの学習及び評価を行った結果である。表8は503文全文で学習し、503文全文で評価を行った品詞遷移学習データについてクローズな実験結果で、PPが約8%改善されている。また、表9は453文で学習し、残り50文で評価を行った品詞遷移学習データについてオープンな実験結果で、PPが約5%改善されている。両者とも、正解句境界を用いて行った表5、表6に比べて効果は落ちているものの、アクセント句境界を自動抽出した場合でも、提案言語モデルが有効であることが分かる。

また表10、表11は同様の実験を話者MHTの音声から自動抽出されたアクセント句境界に基づいて行った結果である。品詞遷移学習データについてクローズな場合で約9%、オープンな場合で約5%のPP改善がみられる。これは表8、表9の話者MYIの結果とほぼ同じ結果であり、提案手法が話者に依存せずに同等の効果をあげていることが分かる。

なお、このように形態素境界付近で自動抽出された句境界に基づいて提案言語モデルのパープレキシティを計算することは、実際の連続音声認識時に正解パスを通ったときの言語尤度がどのようになるのかを調べていることにほぼ相当すると考えられる。従って、表8から表11でPPが改善されたことで、少なくとも正解パスを探索するときには、提案言語モデルによって認識時に言語尤度が有利に働くと考えられる。しかし、正解以外のパスを探索中に、特に句境界の自動抽出の挿入誤りがどのような影響を与えるかについては全く考慮されていないため、これらの結果がそのまま認識率の向上に結びつくかどうかは分からず、今後検討していく必要がある。

3.3.3　話者オープンな場合

表12、表13は、話者MYIの音声から自動抽出された句境界に基づいて学習した提案言語モデルによっ

表8　MYI自動抽出503文で学習及び評価

	全文	句内部	句境界
ベース言語モデル	117.0	57.13	1344
提案言語モデル	107.4	53.75	1133
改善率	8.24%	5.92%	15.7%
バイグラムヒット率	95.35%	96.42%	91.68%

表9　MYI自動抽出句境界453文で学習、MYI自動抽出句境界50文で評価

	全文	句内部	句境界
ベース言語モデル	117.4	57.32	1436
提案言語モデル	111.7	55.12	1331
改善率	4.84%	3.84%	7.26%

表10　MHT自動抽出503文で学習及び評価

	全文	句内部	句境界
ベース言語モデル	117.0	46.40	1932
提案言語モデル	106.9	43.39	1642
改善率	8.63%	6.49%	15.0%
バイグラムヒット率	95.35%	96.73%	91.14%

表11　MHT自動抽出句境界453文で学習、MHT自動抽出句境界50文で評価

	全文	句内部	句境界
ベース言語モデル	117.4	46.54	2082
提案言語モデル	111.4	44.60	1921
改善率	5.07%	4.17%	7.74%

表12　MYI自動抽出句境界503文で学習、MHT自動抽出句境界503文で評価

	全文	句内部	句境界
ベース言語モデル	117.0	46.40	1932
提案言語モデル	109.0	44.04	1701
改善率	6.84%	5.09%	12.0%
バイグラムヒット率	95.35%	96.73%	91.14%

表13　MYI自動抽出句境界453文で学習、MHT自動抽出句境界50文で評価

	全文	句内部	句境界
ベース言語モデル	117.4	46.54	2082
提案言語モデル	110.4	44.41	1893
改善率	5.96%	4.58%	9.10%

表14　MHT自動抽出句境界503文で学習、MYI自動抽出句境界503文で評価

	全文	句内部	句境界
ベース言語モデル	117.0	57.13	1344
提案言語モデル	110.9	54.59	1242
改善率	5.18%	4.45%	7.57%
バイグラムヒット率	95.35%	96.42%	91.68%

表15　MHT自動抽出句境界453文で学習、MYI自動抽出句境界50文で評価

	全文	句内部	句境界
ベース言語モデル	117.4	57.32	1436
提案言語モデル	112.8	55.34	1364
改善率	3.93%	3.45%	5.02%

て、話者 MHT の音声から自動抽出された句境界に基づいたデータを評価した、話者オープンな実験である。表 12 が品詞遷移学習データについてクローズな場合で約 7%、表 13 が品詞遷移学習データについてオープンな場合で約 6%の PP 改善がみられた。

また表 14、表 15 は、同様に話者 MHT で学習し話者 MYI で評価した結果である。表 14 が品詞遷移学習データについてクローズな場合で約 5%、表 15 がオープンな場合で約 4%、PP が改善されている。

これらの結果を表 8 から表 11 と比べると、やはり話者オープンにしたことで全体的に結果が悪くなっているが、表 11 と表 13 のように、話者オープンにしても結果が悪くならない例もみられる。

3.3.4 正解句境界で学習したモデルの利用

表 16 は、話者 MYI の正解句境界に基づいて学習した提案言語モデルで、話者 MYI の音声から自動抽出された句境界に基づいたデータの評価を行った結果である。品詞遷移学習のデータについてクローズな実験条件であるにもかかわらず、提案手法によって PP は悪化している。また、表 17 は同様に正解の句境界に基づいて学習し、話者 MHT の音声から自動抽出された句境界に基づいて評価を行った結果であるが、やはり PP が悪化している。

これは、提案言語モデルの学習時と評価時とで句境界の基準が異なることが原因であると考えられる。従って提案言語モデルを学習するときには、句境界の抽出手法として実際の認識時に利用する句境界の抽出手法を選択することが重要であることが分かる。

表 16 MYI 正解句境界 503 文で学習、
MYI 自動抽出句境界 503 文で評価

	全文	句内部	句境界
ベース言語モデル	117.0	57.13	1344
提案言語モデル	141.2	69.18	1601
改善率	-20.7%	-21.1%	-19.2%
バイグラムヒット率	95.35%	96.42%	91.68%

表 17 MYI 正解句境界 503 文で学習、
MHT 自動抽出句境界 503 文で評価

	全文	句内部	句境界
ベース言語モデル	117.0	46.40	1932
提案言語モデル	131.7	53.33	2041
改善率	-12.6%	-14.9%	-5.66%
バイグラムヒット率	95.35%	96.73%	91.14%

3.4 ベース言語モデルの精度を上げた場合

これまで述べてきた実験はベースの言語モデルとして、毎日新聞記事 1 年分から学習されたバイグラムを利用していた。これに対して表 18 は、ベースの言語モデルとして毎日新聞記事 6 年分から学習された語彙数 20K のバイグラムを用いたときの結果である。正解句境界を用いて学習及び評価を行い、品詞遷移学習データについてクローズな実験を行った。

PP は約 13%改善されており、新聞記事 1 年分から構築されたバイグラムをベースとした表 5 よりも良い結果となっている。これは、ベースの言語モデルの学習量が増えたことで、バイグラムカウントの値が大きくなり、分配がより正確に行われた結果ではないかと考えられる。

表 18 新聞記事 6 年分によるベース LM、
MYI 正解句境界 503 文で学習・評価

	全文	句内部	句境界
ベース言語モデル	109.8	24.28	2444.46
提案言語モデル	95.94	21.85	2013.61
改善率	12.6%	10.0%	17.6%
バイグラムヒット率	97.52%	98.95%	94.53%

4. ま と め

本稿では、アクセント句境界の有無に応じた 2 種類の言語モデルを構築し、これらを句境界の有無によって使い分ける手法を提案した。バイグラム言語モデルを使ってパープレキシティによる評価を行ったところ、正解の句境界、自動抽出された句境界の両方で効果があり、話者オープンな条件でもパープレキシティが改善された。

今後の研究予定としては、まず提案言語モデルを連続音声認識に組み込んで、認識率が改善されるかどうかを検証する。さらに、より大規模な音声データベースを用いて提案言語モデルの学習を行うことで、品詞遷移学習の精度向上を目指す予定である。

文 献

[1] 高橋敏, 松永昭一, 嵯峨山茂樹. ピッチパタン情報を考慮した単語音声認識. 電子情報通信学会技術研究報告 SP90-17, pp. 65–72, 1990.

[2] 山下洋一, 岩橋大輔, 溝口理一郎. 基本周波数パターンを利用したキーワードスポッティング. 電子情報通信学会論文誌, Vol. J81-DII, No. 6, pp. 1065–1073, 1998.

[3] Shi-wook Lee, Keikichi Hirose, and Nobuaki Minematsu. Efficient search strategy in large vocabulary continuous speech recognition using prosodic boundary information. *Proc.ICSLP2000*, Vol. 4, pp. 274–277, 2000.

[4] 阿部匡伸, 匂坂芳典, 梅田哲夫, 桑原尚夫. 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 自動翻訳電話研究所, 1990.

[5] Keikichi Hirose and Koji Iwano. Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. *Proc.ICASSP2000*, Vol. 3, pp. 1763–1766, 2000.