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

2004－SLP－54（16）
2004／12／21

頑健な区間検出とモデル適応に基づく雑音下音声認識

張 志鵬† 古井 貞熙††

†ＮＴＴドコモ　マルチメディア研究所

〒239－8536　神奈川県横須賀市光の丘3－5

††東京工業大学大学院 情報理工学研究科　計算工学専攻

〒152-8552 東京都目黒区大岡山 2-12-1

E-mail: †zzp@mml.yrp.nttdocomo.co.jp, ††furui@cs.titech.ac.jp

あらまし　本論文では、雑音やSNRが時間的に変化する状況に対処することを目的として、頑健な区間検出とモデル適応に基づく雑音下音声認識手法を提案する。入力音声に対し、一定の長さの入力信号を切り出し、その区間に対して木構造から選ばれた最適な雑音適応モデルによって音声認識する。その結果から文区切りを検出し、さらに尤度が最大化するように音素モデルを教師なしで線形変換（MLLR）を行ったのち、再度音声認識を行うことによって認識精度を上げる。日本語対話システムを用いて、二種類の雑音データに対する評価実験により提案手法の有効性を確認した。

キーワード　時変雑音，　音声区間検出，　モデル選択，　雑音適応法

Noisy Speech Recognition Based on Robust End-point Detection and Model Adaptation

Zhipeng ZHANG† and Sadaoki FURUI††

† NTT DoCoMo, Inc. Multimedia Laboratories

3-5 Hikarinooka, Yokosuka-shi, 239-8536 Japan

††Tokyo Institute of Technology, Department of Computer Science

2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan

E-mail: †zzp@mml.yrp.nttdocomo.co.jp, ††furui@cs.titech.ac.jp

Abstract How to detect speech periods in noisy speech and how to cope with the temporal variation of noise characteristics are challenging problems. This paper proposes a new robust noisy speech recognition method based on robust end-point detection and online model adaptation using tree-structured noisy speech HMMs. The basic algorithm consists of 1) blind speech segmentation, 2) best matching GMM selection, 3) recognizing the speech with the HMM that corresponds to the GMM, 4) end-point detection based on the recognition results, 5) HMM adaptation based on the recognition results, and 6) re-recognition using the adapted HMM. The processes of 1) through 6) are repeated by shifting the blind segmentation window until the end of the sequence of utterances is detected. The proposed method is evaluated by noisy speech collected by a Japanese dialogue system. Experimental results show that the proposed method is effective in recognizing noisy speech under various noise conditions.

Keywords: Time-variable noise, End-point detection, Model selection, Noise adaptation

1. はじめに

音声認識実験の多くは、入力音声の端点が既知との仮定に基づいて行われるが、実環境における入力音声は端点が未知の連続入力ストリームである。端点が未知の音声、特に雑音条件下の連続入力音声を自動的に認識する技術が必要である。

雑音条件下の音声認識の難点は雑音特性とSNRが時間的に変化することである。本論文では、そのような状況において、音声区間を自動的かつ頑健に検出し、さらに音声認識精度を向上させるために、音素モデルをオンラインで雑音に適応化する方法について検討する。

文の区切り情報がない連続入力音声を認識する技術は、二種類に分類できる。ひとつは直列処理手法である[1,2]。まず信号処理技術を用いて音声セグメンテーションを行い、次に音声認識を行う方法である。もうひとつの手法は音声のセグメンテーションと認識を同時に行う並列処理手法である[3,4]。この手法は音声と非音声のモデルを学習しておき、入力の連続音声に対しこれらのモデルを用いて認識を行う。認識結果の中の音声と非音声の区切りによってセグメンテーションが決まる。以上の二種類の従来の手法は、雑音の特徴が既知であるか、定常であれば高い性能を得ることができるが、雑音やSNRが変化する環境においては性能が低下する。

我々はこれまでに、木構造クラスタリングに基づく区分線形変換法による雑音適応法を提案している[5]。この方法では雑音特性を階層的に逐次分割することにより、雑音重畳音声モデルの木構造を作成する。木構造で雑音特性を表すことにより、木構造の上層では雑音特性の大局的な特徴、下層では特定の雑音特徴を表現するモデルが得られる。この木構造を上から下にたどり最適なモデルを選択することにより、最適な雑音区分空間を選択できる。さらに選ばれるモデルに対し尤度が最大となるように線形変換を行い、これを用いて音声を認識する。種々の雑音条件での雑音重畳音声に対して、提案手法の効果が確認されている。

本論文では、木構造クラスタリングに基づく区分線形変換法を用いた、頑健な区間検出とモデル適応に基づく、雑音下音声認識手法を提案する。入力音声に対し木構造から選ばれた最適な雑音適応モデルによって音声認識し、その結果から文区切りを検出し、さらに尤度が最大化するようにモデルの線形変換[6] (MLLR) を行って再認識することにより、認識精度を上げる。

2. 区分線形変換に基づく雑音下音声認識

図1で提案手法の構成を示す。本手法は学習段階のモデル作成(言語モデルと木構造雑音重畳音声モデル)と、連続音声認識の二段階からなる。

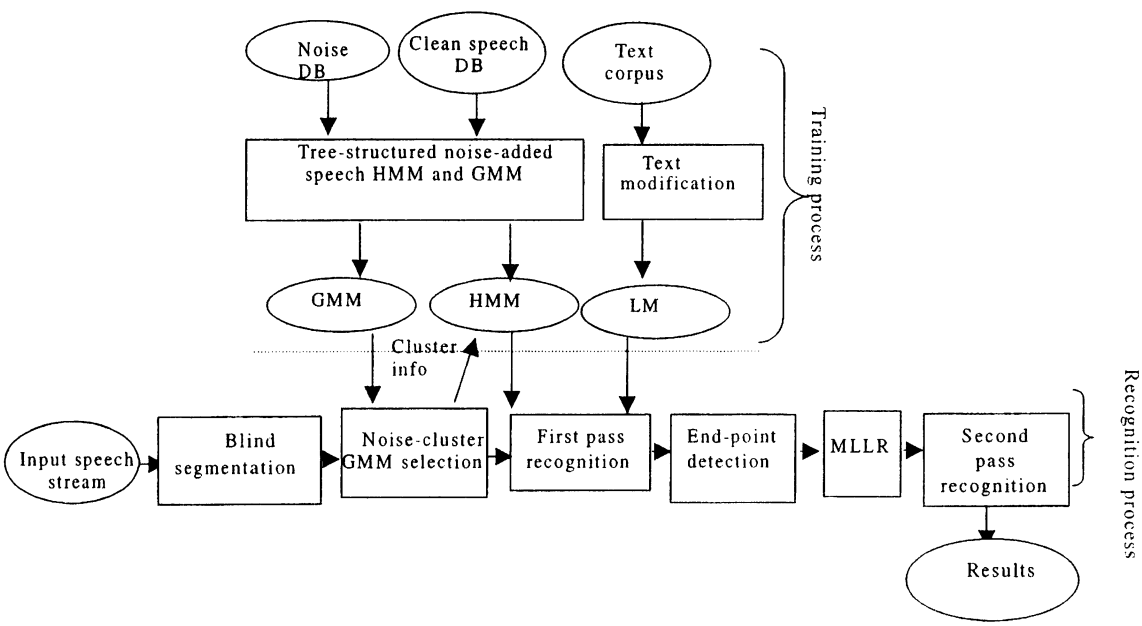

Fig. 1. System flow of the proposed method.

2.1. 文区切りを含めない言語モデル

文の区切りがない連続入力の音声を認識するために、対応する言語モデルを構築する。各文に始点と終点記号があるすべての学習用のテキストデータを接続し、一文にまとめる。前の文の終点記号と次の文の始点記号を一つのマークLP(long pause)に変更する。このように修正されたデータを用いて、言語モデルを学習する。同時に、このLPモデルに対応する音響モデルも用意する。

2.2. 木構造雑音重畳音声モデルの作成

多様な種類の雑音とSNR条件下の雑音重畳音声モデルを学習し、雑音重畳音声のクラスタリングを行って、一つの木構造を作成する。この木構造のルートモデルはすべてのSNR条件のあらゆる雑音重畳音声を含み、葉ノードは特定SNR条件の一つの雑音重畳音声データから学習したモデルを表す。この木構造をルートからリーフ方向にたどり、最適なモデルを選択することにより、入力音声に最適な雑音区分空間を選択できる。

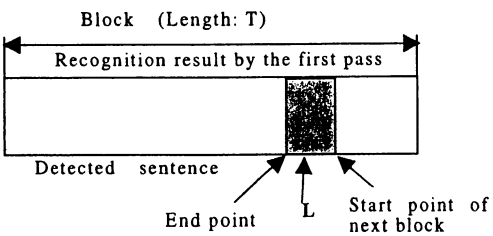

Fig. 2. End point detection
(Case 1: LP is found in the recognition result)

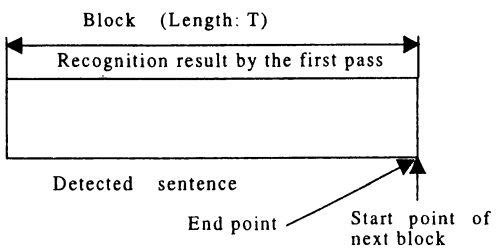

Fig. 3. End point detection
(Case 2: LP model is not found in the recognition result)

2.3. 連続音声認識

連続入力音声には、明示的に文の区切り情報が与えられないため、まず固定長の音声(ブロック)を抽出し、ブロックごとに処理する。雑音条件の変化に対応できるようにするため、以下のようにモデル選択、一回目の認識、セグメンテーション、モデル適応、再認識処理を行う。前のブロックを処理したら、次の固定長の音声に対し同じ処理を行う。その結果、比較的短い遅延で連続音声認識を行うことができる。

1) ブラインドセグメンテーション

連続入力の音声から長さTの音声をブロックとして抽出し、以下のステップ2)からステップ6)までの処理を行う。4)の処理で検出された文の終端を手掛かりに、新たに次の長さTのブロックを抽出し、それに対して、同じ処理2)～6)を行う。

2) GMMを用いた最適モデルの選択

各入力ブロックに対し、木構造空間を上から下にたどり最適なモデルを選択することにより、入力音声に最適な雑音区分空間を選択できる。

多数のモデルから尤度最大のHMMを選択するのは計算量が膨大になるため、クラスタ雑音重畳HMMの代りに、クラスタ雑音重畳GMMを用いて選択する手法を用いる。クラスタ雑音重畳GMMはクラスタ雑音重畳HMMの学習に使われるデータと同じものを用いて作成する。入力音声に対し、最大尤度を示すクラスタ雑音重畳GMMに対応するHMMを選択する。

3) 一回目の認識

選択されたモデルと2.2節で説明した言語モデルを用い、音声認識を行う。

4) 端点検出

認識結果に基づいて以下のように端点検出を行う：認識結果の中にLPモデルがある場合は、最初のLPモデルの出現場所を端点とする。認識結果の中にLPモデルがない場合は、この音声ブロックの終点を端点とする。図2と図3に、LPモデルが検出される場合と、されない場合のプロセスを示す。

5) モデル適応

高精度の認識性能を達成するために、選ばれたモデルに対しMLLRによるモデル適応を行う。MLLR変換手法は、HMMのガウス分布の平均値及び分散を尤度最大化の規準に基づいた線形変換により適応化する方法である。本研究では、極めて短い音声区間を用いて適応を行うので、すべての音素を一つの共通の変換行列を用いて適応

化する。

6) 適応化モデルによる再認識

MLLR 変換を行ったモデルを用い、再度音声認識を行って、最終結果を出力する。

3. 実験条件

3.1. 音響モデル

実験で利用する音響モデルは[話し言葉工学]プロジェクト[7]で作成された音声コーパス中の男性話者による 338 講演（約 59 時間分）の音声データを用いて作成した、2,000 状態 16 混合の状態共有型 triphone HMM である。音声特徴量は、MFCC および MFCC とパワーの動的特徴からなる 25 次元のベクトル系列を用いた。

3.2. 言語モデル

音声認識実験のタスクとしては、音声入力による飲食店舗検索の対話システムを用いた。ユーザーは場所(最寄り駅)、料理の種類などの検索条件を発声することにより、希望の店舗の詳細情報を得る。発話内容を受理する言語モデルは、それぞれの発話内容用のテキストコーパスから作成される。コーパスの量の不足に対処するため、クラス言語モデルを用い、クラスに属する単語は全て等確率で生起すると仮定した[8]。

3.3. 学習用雑音データ

学習用雑音データは電子協雑音データベースの 28 種類の雑音を用いた。

3.4. 評価用データ

2 種類の評価用データを用いた。

- **Test-1**：10 名の話者が発声した計 50 発話の対話音声に対し、3 種類の SNR (SNR=5,10,15dB)で、電子協の雑音のうち学習に用いなかった 2 種類の雑音（"station"、"hall"）を、計 6 種類の組合せで重畳させたデータを用意した。
- **Test-2:** 3 日間にわたって 2 箇所（横浜駅" station"、オフィス" office"）の実際の雑音環境で収録した 12 名の話者の 540 文(平均 SNR は、それぞれ 10dB と 12dB)を使用した。

4. Test-1 における実験結果

4.1. 全般的な結果

まず、Test-1 の雑音重畳音声に対し以下の 3 種類の条件下で音声実験を行った。

a. Baseline:クリーン HMM を用いて認識を行う

b. 提案法(Proposed method):提案法により端点検出を行い、モデル適応を経て認識を行う

c .区切り指定(Given end-point):文の正しい区切り情報を指定した場合(指定した端点情報に基づき木構造から最適なモデルを選択しさらにモデル適応を行う)

これらの実験においては、切り出す音声ブロックの長さを 10 秒に設定した。2 種類の雑音に関して、SNR を 5, 10, 15dB に変えた場合の単語正解率(ACC%)を、表 1 と 2 に示す。提案手法は、いずれの雑音の場合も、Baseline に比べて、大幅によい性能を示している。また区切り指定の場合に近い性能が得られることがわかる。

Table 1. Recognition accuracies (%) with Test 1 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method (Station noise-added speech)

SNR	Baseline	Given end-point	Proposed method
5dB	23.7	40.8	38.2
10dB	57.9	67.1	67.1
15dB	67.1	76.3	73.7

Table 2. Recognition accuracies (%) with Test 1 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method (Exhibition hall noise-added speech)

SNR	Baseline	Given end-point	Proposed method
5dB	40.8	47.4	43.4
10dB	61.8	76.3	75.0
15dB	71.1	77.6	77.6

4.2. ブロック長の学習用雑音データ

Test-1 の雑音重畳音声に対し、ブラインドセグメンテーションのブロック長(T)の影響を調べた。T が 3,5,8,10 秒の場合の比較実験を行った。

2 種類の雑音に関して、SNR を 5, 10, 15dB の 3 種類に変えた場合の単語正解率 (ACC%)を、図 4 と図 5 に示す。T が 5 秒以上の場合は性能がほぼ変わらず、T が 3 秒の場合には性能が落ちる。この長さの閾値は、入力音声の最大の長さにほぼ一致している。

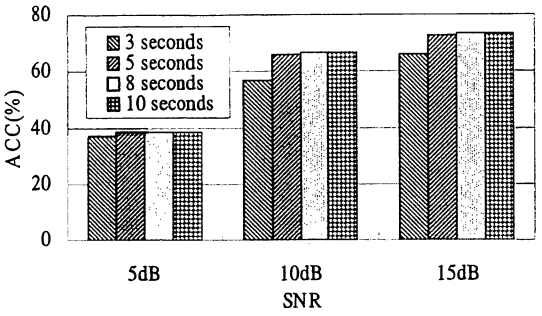

Fig. 4. Recognition accuracies (%) on Test 1 for various blind segmentation lengths (Station noise-added speech)

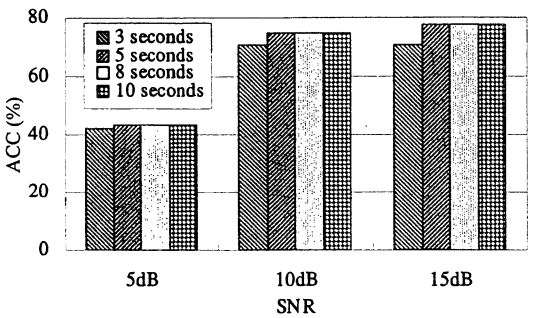

Fig. 5. Recognition accuracies (%) on Test 1 for various blind segmentation lengths (Exhibition hall noise-added speech)

4.3. MLLR の効果

次に Test-1 の雑音重畳音声に対し、MLLR の効果を調べるため、モデル選択のみを行い、選択されたモデルへの MLLR は適用しない場合の実験を行った。この実験において音声ブロックの長さは 10 秒に設定した。表 3 と表 4 に MLLR を実施しない場合の単語正解率 (ACC%)を示す。表 1 と表 2 の結果に比べて顕著に性能が低下しており、モデル選択の後に MLLR を行うことが有効であることがわかる。

Table 3. Recognition accuracies (%) with Test 1 data when MLLR adaptation is not applied (Station noise-added speech)

SNR	Baseline	Given end-point	Proposed method
5dB	15.8	36.8	35.5
10dB	47.4	64.7	60.5
15dB	63.2	75.0	65.8

Table 4. Recognition accuracies (%) with Test 1 data when MLLR adaptation is not applied (Exhibition hall noise-added speech)

SNR	Baseline	Given end-point	Proposed method
5dB	30.3	47.4	40.8
10dB	51.3	70.3	69.7
15dB	67.1	75.0	72.4

5. Test-2 における実験結果

つぎに Test-2 の雑音重畳音声に対し、T を 10 秒に設定したの場合の実験を行った。表 5 にクリーン HMM を用いた場合 (Baseline)、提案法 (Proposed method)、区切り指定(Given end-point) の 3 種類の条件下の実験結果を示す。この結果から、Test-2 に対しても、提案手法は Baseline よりかなり性能がよいことが分かる。また区切り指定の場合に近い性能を得られることが確認された。

Table 5. Recognition accuracies (%) with Test 2 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method

Noise	Baseline	Given end-point	Proposed method
Station noise added speech	40.8	56.6	55.8
Office noise added speech	55.2	73.4	72.5

6. まとめ

頑健な区間検出とモデル適応に基づく雑音下音声認識手法を提案した。雑音が重畳した入力音声を、一定の長さのブロックで抽出 (ブラインドセグメンテーション) し、モデル選択、音声区間検出、モデル適応、再認識処理を行う。日本語対話システムにおいて、2 種類の雑音データに対する評価実験により、提案手法の有効性が確認された。本提案手法は、雑音重畳入力音声に対して、オンラインで処理を行うことができるが、 [ブロック長＋再認識を含む処理時間]の遅延が生ずる。ブロック長としては、入力音声の長さに依存するが、本実験では、5 秒長以上あれば十分であることが確認された。実時間性が強く要求される場合には、モデル選択および適応の処理の回数を適宜間引くことによって、遅延をいくらか減らすことは可能である。

現在の処理法では各ブロックに対しブラインドセグメンテーション・モデル選択・適応を行うため、少なくとも 1 ブロックの長さの遅延が生じ

る。処理時間を削減し実時間で動作するようにするには、現在のブロック処理に前のブロックに最適なモデルを用いて、適応化と認識処理を並列に行うことにより、この遅延が削減できる。一個先のブロックに対する最適なモデルの選択と、現在のブロックに対するセグメンテーションの処理は、複数の CPU による並列処理を用いることで可能になる。今後の課題には、以上のような実時間化に向けた改善を含めて、より実用性に優れた方法への改良や、学習に用いる雑音の種類の増加などがある。

文　献

[1] L.R.Rabiner : "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical Journal, pp. 297-315, 1975

[2] 新美, "音声認識",共立出版社, 1979

[3] A. Acero: "Robust HMM-based end-point detector", Proc. Eurospeech, pp. 1551-1554, 1993

[4] J.G.Wilpon et al.: "Application of hidden Markov model to automatic speech end-point detection", Computer Speech and Language, pp. 321-341, 1987

[5] Z.P. Zhang et al.: "A tree-structured clustering method integrating noise and SNR for piecewise-linear transformation-based noise adaptation", Proc. ICASSP, pp. 981-984, 2004

[6] C.J.Leggetter et al., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". Computer Speech and Language, Vol.9, pp.171-185, 1995.

[7] 古井, 他, "科学技術振興調整費開放的融合研究推進制度　－大規模コーパスに基づく「話し言葉工学」の構築－", 日本音響学会誌, vol.56, no.11, pp.752-755, 2000

[8] 田熊, 他, "並列処理型計算機による混合主導型音声対話システムの構築", 秋季音講論, pp.79-80, 2002