T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	区分線形変換による雑音適応法における木構造クラスタリングの効果
Title(English)	
著者(和文)	張 志鵬, 大辻 清太, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 2002年秋季講演論文集, Vol. , No. 1-9-15, pp. 29-30
Citation(English)	, Vol. , No. 1-9-15, pp. 29-30
発行日 / Pub. date	2002, 9

Powered by T2R2 (Science Tokyo Research Repository)

区分線形変換による雑音適応法における木構造クラスタリングの効果*

◎張 志鵬　大辻 清太 (NTTドコモ) 古井 貞熙 (東工大)

1．はじめに

大語彙連続音声認識における問題の一つとして，背景に雑音や音楽を含む音声に対する認識性能の劣化が挙げられる。これまでに我々は区分線形変換(PLT; piecewise linear transformation)による雑音適応法[1]を提案した。この論文では区分線形変換における雑音の木構造クラスタリングの効果について考察する。

2．区分線形変換雑音適応法における木構造クラスタリング

2.1 区分線形変換による雑音適応手法

一般に、音声に雑音が重畳されたときに、雑音音声信号 $\hat{s}$ は次のようにモデル化される。

$$\hat{s} = F(s, n, SNR)$$

s ,n,SNRはそれぞれクリーンな音声信号、重畳する雑音、信号対雑音比を表す。Fはケプストラム空間では一般に非線形変換になる。この問題に対応するために、ケプストラム空間での確率分布を表すHMMに対して、HMM合成法[2]やneural network法[3]などの種々の非線形処理が研究されてきた。しかし、これらの手法には二つの欠点がある。①複雑な処理と大きな計算量を必要とする。②雑音n及び信号対雑音比SNRが常に変動するため、各入力文ごとに適応化する場合には、大きな問題となる。これに対し、我々は非線形処理を区分線形変換で近似して,モデルの尤度最大化をはかる方法を提案した[1]。HMMパラメータ空間（雑音が重畳した音声のHMM空間）を雑音の性質とSNRによって区分化し、入力音声の条件に最も適合した部分空間を選ぶ。選ばれた空間で、尤度がさらに最大化するように線形変換（MLLR[4]）を行う。

2.2 木構造クラスタリング

以前の手法[1]ではクラスタ数の決定はヒューリスティックに行った。一般にクラスタ数の設定の問題を考えるとき、本来いくつのクラス数が最適であるかは不明である。クラス数が異なると一つのクラスがもつ分散の大きさが異なってくる。データが少ない場合や学習データと入力のミスマッチが大きいときはクラス数を少なく、データ数が多い場合や学習データと入力のミスマッチが小さいときはクラス数を多くする方がよいと考えられる。雑音クラスタリングに最適なクラス数を決めるには何らかの基準が必要である。そこで本報告では階層的な雑音クラスタリングによる雑音適応法を提案する。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、下層では局所的な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。概念図を図1に示す。まずSNRでクラスタ化し、次にSNR条件ごとに木構造を作成する。認識するときは、あらゆるSNRの条件から尤度最大なノードを選択する。選ばれた空間で、尤度がさらに最大化するように線形変換（MLLR）を行う。

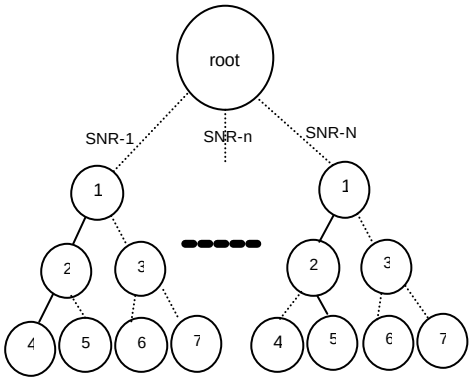

Fig. 1: Tree-structure of noise clustering

3．認識実験

3.1 音響モデル

音声HMMとしてtree-based clusteringにより状態共有化を行った不特定話者文脈依存音素HMMを用いる。音響特徴量としては16次のLPCケプストラムと対数パワー,及びそれらの一次微分の計34次元を使用した.学習用クリーン音声データは,男性53名による13,270発話である。モデルの総状態数は2,106，各状態のガウス分布の混合数はすべて4である。

3.2 言語モデル

言語モデルの学習に用いたデータは放送ニュース原稿テキスト5年分、約50万文である。単語出現頻度上位2万語を認識語彙とし、間投詞を考慮した言語モデル[5]を用いた。

3.3 学習用雑音データ

雑音データは電子協雑音データベースの30種類及びNTTドコモが収録した8種類計38種類の雑音を用いた。Baum-Welchアルゴリズムを用いて64混合の各雑音GMMを学習した。

---

* A tree structure noise clustering method for piecewise linear transformation-based noise adaptation
By Zhipeng Zhang, Kiyotaka Otsuji (NTT DoCoMo) and Sadaoki Furui (Tokyo Institute of Technology)

3.4 評価用データ

2種類の評価用データを用いた。まず、1996年7月に実際に放送されたニュース音声から,背景に多種の雑音や音楽が乗っている発話や記者レポートなどの発話50文(Test1、平均SNR=17dB)を使用した。またNTTドコモが収録した3種類の雑音環境(社内オフィス“Office”、大通り“Street”、横浜地下街“Shop”)で、収録した4名の男性話者による100単語(1人25単語),計300単語のデータを用意した(Test2、平均SNR=24,17,18dB)。

4．認識実験結果

Test1に対し、HMMモデルのSNR=10dBの場合の実験を行った。この木構造を尤度最大基準に基づいて、ルートから下にたどり最適なノードを選択し、尤度がさらに最大化するように線形変換を行う。認識正解精度(Acc%)を“tree”で図2に示す。雑音クラスタ数を2,4,8,16,24,38に固定した時の比較実験の結果も図2に示す。各クラスタ数の条件の内、認識率が最高の24クラスタよりも木構造クラスタリングの方が認識精度が高い。

また、HMMモデルのSNR=15dBの場合の実験結果を図3に示す。認識率が最高の8クラスタよりも木構造クラスタリングの方が認識精度が高い。次に、50文ごとに木構造の各階層における選択結果を調べた。その結果、最上層のモデルが選択された文が37あった。このことから、実際放送される音声の雑音環境は学習に用いられる雑音と大幅に違うことが考えられる。最上層以外のモデルが選択された13文に関しては、木構造の真中付近のモデルを選択している。これは学習に用いた雑音だけではすべての雑音空間を埋めるには不足し、入力雑音音声は複数の雑音を組み合わせたモデルに近い場合が多いことを示している。

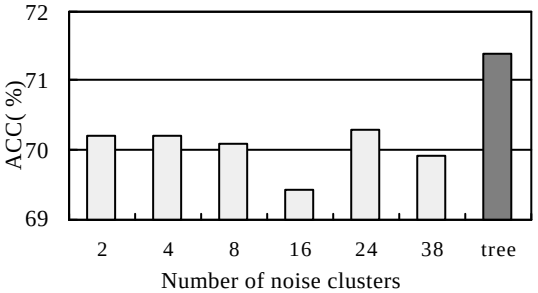

Fig. 2: Recognition results for Test1 (Model SNR=10dB)

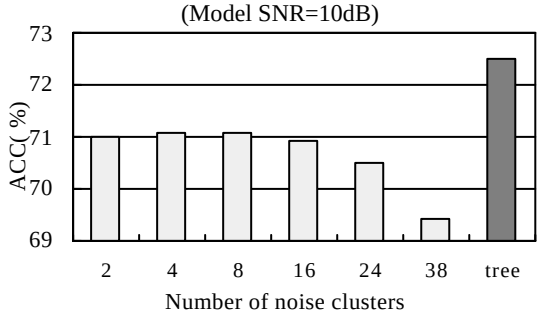

Fig. 3: Recognition results for Test1 (Model SNR=15dB)

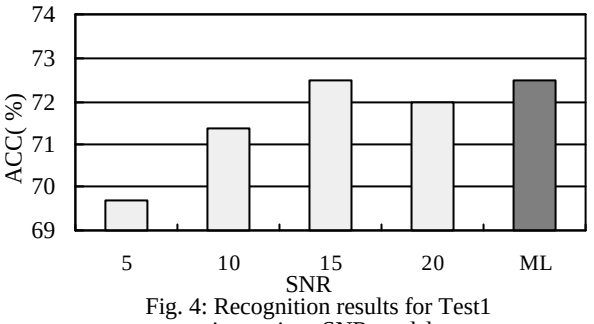

Fig. 4: Recognition results for Test1 using various SNR models

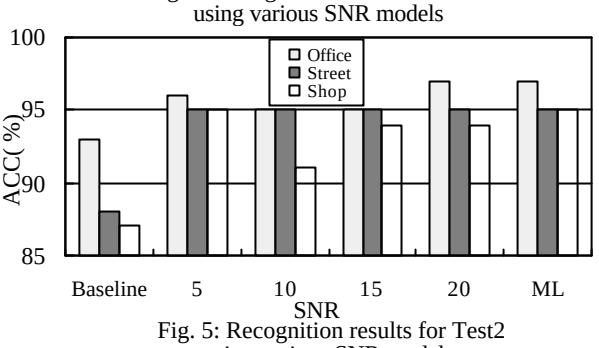

Fig. 5: Recognition results for Test2 using various SNR models

図4に各SNR(5,10,15,20dB)の木構造モデルから最尤のノードを選択する場合の実験結果を示す。入力音声の平均SNRに一番近いSNRが15dBの場合が一番良い結果になることが分かる。あらゆるSNR条件の木構造モデルから尤度最大のモデルを選択する場合の実験結果“ML”を同じ図に示す。この場合、ベースラインに比べ単語誤り率は26.1%低下した。

次にTest2の3種類の雑音環境に対し適応実験を行った。各SNR(5,10,15,20dB)の場合の木を用いた場合の実験結果を図5に示す。入力音声の平均に一番近いSNRの場合(SNR=20dB)が一番良い結果になる。また、あらゆるSNR条件の木構造モデルから尤度最大のモデルを選択する場合の実験結果“ML”も同じ図に示す。ベースラインに比べ、三種類のデータの平均で単語誤り率は59.3%低下した。

5．まとめ

尤度最大化規準に基づく区分線形雑音適応法における雑音の木構造クラスタリングについて検討した。二種類の実環境での雑音重畳音声に対して、提案手法の効果を確認した。今後の課題には、学習雑音数の増加、木構造クラスタリングの改善、MDLによるクラスタの選択、雑音区間の自動切り出しなどがある。

謝辞

本研究は、東工大・古井研究室で行われたものである。討論いただいた研究室の方々に感謝する。ニュース原稿及び音声データを提供して頂いたNHK放送技術研究所に感謝します。

参考文献

[1]張, 古井, 秋季音講論, pp.29-30, 2001.
[2]F.Martin et al., 信学技報 SP92-96, 1992.
[3]張, 古井, 春季音講論, pp.55-56, 2001.
[4]C.J.Leggetter et al., Computer Speech and Language, Vol.9, pp.171-185, 1995.
[5]桜井 他, 春季音講論, pp.57-58, 1999.