尤度最大化規準による雑音適応

張 志鵬　古井 貞熙
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
{zzp,furui}@furui.cs.titech.ac.jp

尤度最大化規準による音素 HMM の雑音適応を実現するため、非線形処理を区分線形変換で近似する方法を提案する。雑音重畳音声の HMM パラメータ空間を区分化し、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）を行う。また、MLLR における分散の適応効果を考察する。二種類の雑音データ(人工的に付加した雑音音声と,実際に放送されたニュース音声で背景に多種の雑音や音楽が乗っている雑音音声)に対する評価実験によって提案手法の有効性を確認した。

Piecewise-Linear Transformation-based HMM Adaptation for Noisy Speech

Zhipeng Zhang and Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{zzp, furui}@furui.cs.titech.ac.jp

This paper proposes a piecewise-linear transformation method for noise adaptation of phone HMM. Various noises are clustered according to their acoustical property and signal-to-noise ratios (SNRs), and noisy speech HMM is made for each clustered noise condition. Based on the likelihood maximization criterion, the HMM which best matches an input speech is selected and further adapted using the MLLR method. The proposed method was evaluated by recognizing noisy broadcast-news speech. It was confirmed that the proposed method was effective in recognizing numerically noise-added speech as well as actual noisy speech uttered by a wide range of speakers under various noise conditions.

1. はじめに

近年音声認識技術が著しく進展し、数多くのアプリケーションが実現されるようになった。しかし、まだ多くの研究課題が残っている。認識誤りの主たる原因は、学習データと認識すべき音声の間になんらかのずれがあることである。その背景には、話者の個人差が極めて大きいこと、種々の雑音とマイクロホンや伝送系などの歪みなどがあげられる。これらの音響変動に対応する耐性の向上技術が極めて重要である。

雑音に関しては種々の変動に対応できる基本的な適応化法として有望なのが HMM 合成法[1]である。この手法では、音声の HMM と雑音の HMM の畳み込みをスペクトル空間で行って、雑音が重畳した音声の HMM を合成する。しかしながら、HMM 合成法には計算量が多い、CMS と併用できないという問題点があるため、これに対処する方法としてニューラルネットワーク[2]を用いて雑音適応手法を提案した。ニューラルネットワーク法では、HMM の

各状態の出力確率分布に対し、ニューラルネットワークを用いて雑音に適応した値を推定する非線形演算を行う。ニューラルネットワーク法はほとんどの条件でHMM合成法より優れた適応性能を有することがわかった。ニューラルネットワーク学習時と違う種類の雑音にも有効であることが確認できた。

尤度最大化規準が音声認識の各方面、例えばデコーディング、話者適応などによく用いられる。雑音適応に関しても、特に雑音が時間的に変化する場合に尤度最大化によるモデル適応が有効と考えられる。その一つの方法として、拡張HMM合成法[3]が提案された。しかしながら、この手法では大きな計算量を必要とする問題がある。計算量を減らすために、特徴パラメータ領域で、雑音の影響を尤度最大化規準に基づいて推定し、入力音声から除去する手法[4]が提案された。しかしながら非線形な効果に対して、入力信号からバイアスを引くだけでは不十分であるため、この手法には限界がある。そこで本論文では、非線形処理を区分線形変換(PLT; piecewise linear transformation)で近似して,モデルの尤度最大化をはかる一つの方法を提案する。なお、すでに音素クラスタ別にMLLR変換を適用する雑音適応手法[5]が提案されているが、ここで提案する手法は、雑音の特性まで含めて区分化するところに特徴がある。評価には実際に放送されたニュース音声の認識をタスクとして用いた。

2. 区分線形変換による雑音適応手法

提案する方法では、HMMパラメータ空間(雑音が重畳した音声のHMM空間)を区分化し、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)を行う。図1にこの手法の流れを示す。

2.1 雑音クラスタリング

多数の雑音とSNRの条件に対し、HMMパラメータ空間を区分化(クラスタ化)し、各クラスタについて音響モデル(クラスタ雑音重畳HMM)を作成する。入力雑音重畳音声に一番近いクラスタ雑音重畳HMMを選択し、認識に用いる。

多種類の雑音を直接区分化するのは困難なので、各雑音データから作成された雑音GMMにより区分化を行う。雑音GMMに対し尤度を計算し、雑音間の尤度行列に基づいて、SPLIT法[6]で用いられたクラスタリングアルゴリズムを適用した。

クラスタ雑音重畳HMMはクリーン音声モデ

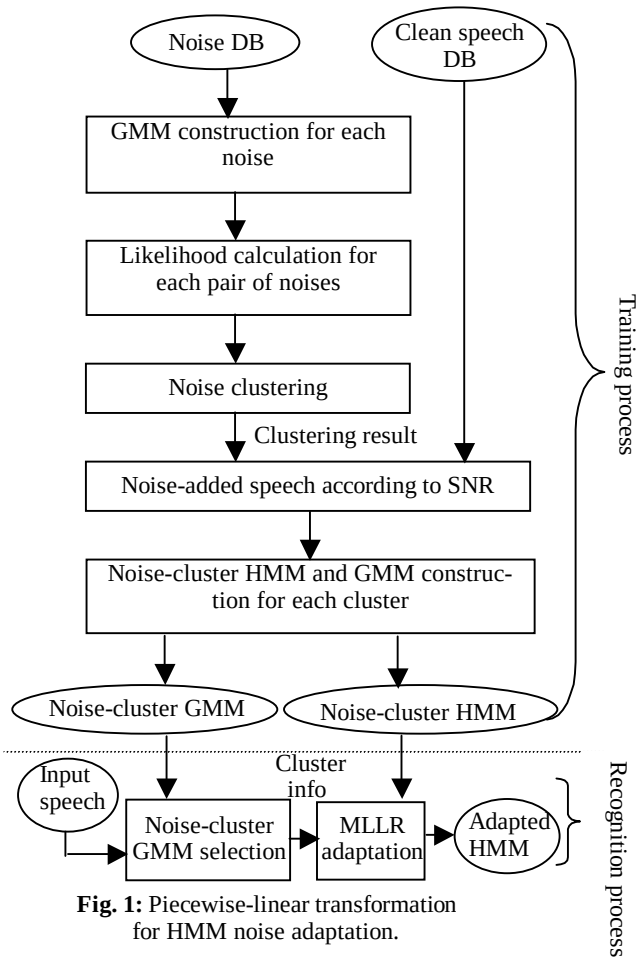

Fig. 1: Piecewise-linear transformation for HMM noise adaptation.

ルを初期モデルとして、クラスタに属する雑音を種々のSNRで重畳した音声データを用いた連結学習により作成する。

2.2 線形変換による雑音適応手法

MLLR[7]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。次の式のようにガウス分布の平均値を更新する。

$$\hat{\mu} = A\mu + b \tag{1}$$

ただし、Aは$n \times n$の行列である。分散をも考慮する場合[7]は変換後の共分散行列は以下のように計算する。

$$\hat{\Sigma} = LHL^T \tag{2}$$

Hは変換行列で、Lは変換前の共分散行列のCholeski分解である。本研究では、各クラスタ雑音重畳モデルを、MLLRを用いて適応化する。また、平均だけと平均分散両方を適応する手法について検討した。

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,ATR音声データベースBセット,日本音響学会連続音声データベース,および同模擬対話データベース中の,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[8]を用いた。

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。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。

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一人の男性話者による10文のクリーンなニュース音声に,3種類のSNR (SNR=0,10,15dB)で、学習に用いなかった2種類の雑音（人ごみと展示場）計6種類の組合せを重畳させたデータを用意した(Test-1)。また、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(Test-2、平均SNR=17dB)を使用した。

4. 認識実験結果

4.1 雑音クラスタリングによる効果

Test-1の6種類の雑音重畳したデータに対し、同じSNRのクラスタ雑音重畳HMMを用いて、クラスタ数を8,16,28とした時のクラスタ選択による適応実験を行った。各SNRにおける単語正解精度を図2,3に示す。全体として16クラスタ程度が最適であることが分かる。

またTest-2に対し各SNRのクラスタ雑音重畳HMMを用いて適応実験を行った。各SNRにおける単語正解精度を図4に示す。入力音声のSNR(17dB)に一番近い15dBのモデルで28クラスタの場合が一番よい結果を示している。

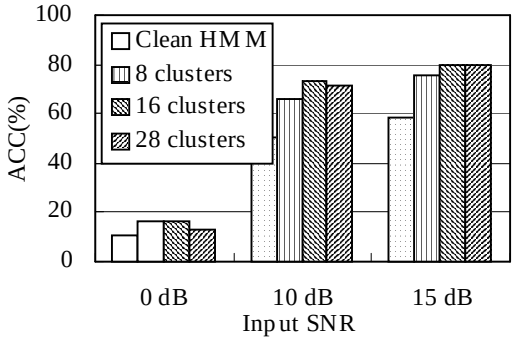

Fig. 2: Recognition results by noise-cluster HMM selection for Test-1 (crowd noise-added speech).

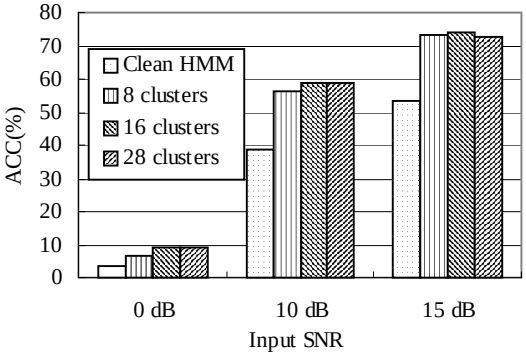

Fig. 3: Recognition results by noise-cluster HMM selection for Test-1 (exhibition hall noise-added speech).

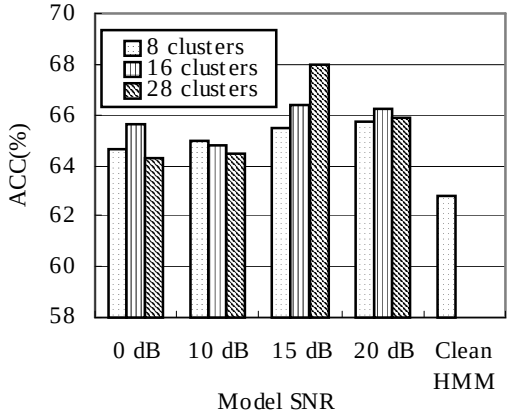

Fig. 4: Recognition results by noise-cluster HMM selection for Test-2.

4.2 区分線形変換による適応結果

0,10,15,20dB の SNR におけるあらゆる雑音音声モデルの中から最適なクラスタ雑音重畳 HMM を選択し、さらに線形変換手法を用いてモデル平均値の適応を行う。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

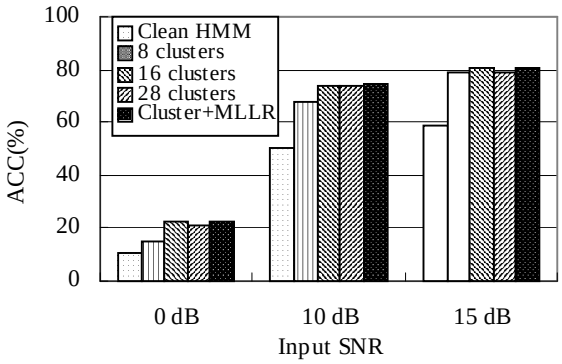

Fig. 5: Recognition results using PLT for Test-1 (crowd-noise-added speech).

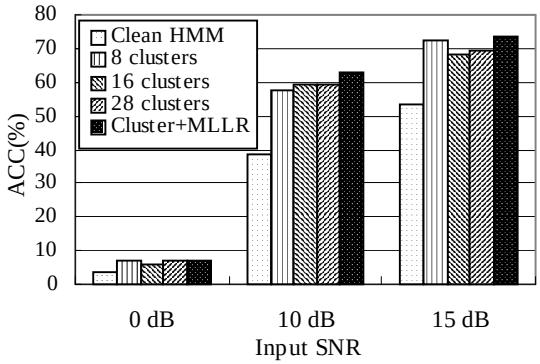

Fig. 6: Recognition results using PLT for Test-1 (exhibition hall noise-added speech).

りにクラスタ雑音重畳 GMM を用いて選択する手法を用いた。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した。

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 を選択し、モデルの平均値の線形変換を行った。あらゆる音素を同じ変換行列によって変換した。図 5,6 に Test-1 の雑音重畳音声に対し適応を行った効果を示す。

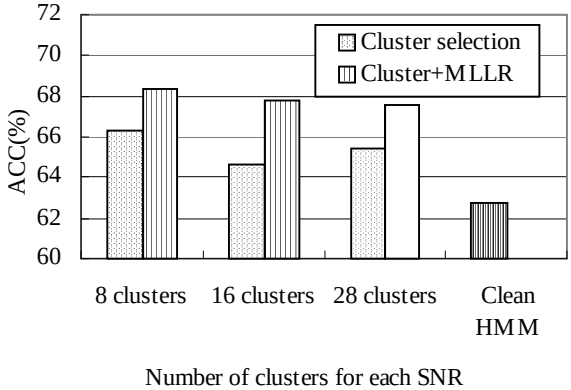

Fig. 7: Recognition results using PLT for Test-2.

また Test-2 に対し適応実験を行った。結果を図 7 に示す。これらの結果から、音声に多種類の雑音が加わり、雑音が変動しているような環境の場合での本手法の有効性が確認された。

4.3 区分線形変換における音素クラスタリングと分散適応の結果

ここまで述べた手法では HMM モデル各分布の平均値だけを線形変換した。また、あらゆる音素を同じ変換行列によって変換した。

次に HMM モデルの音素をクラスタに分けて線形変換を行った。さらに、HMM モデル各分布の平均と分散の両方を適応する実験を行った。図 8,9,10,11 に雑音クラスタ数 16 の場合の Test-1 雑音重畳音声に対し適応を行った効果を示す。雑音の種類によって分散の適応効果がある場合とない場合があることが分かる。効果がある場合でも、その度合いが小さい。音素クラスタの数を増やしても性能は上がらないことが分かる。

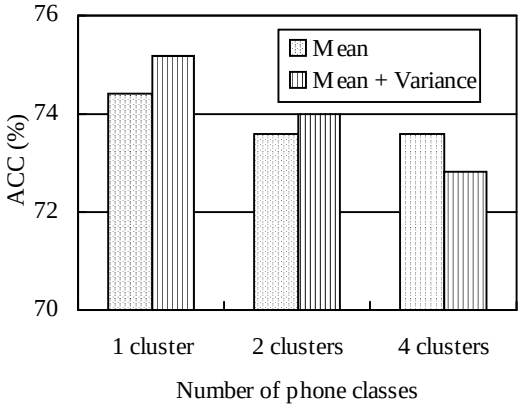

Fig. 8: Recognition results using PLT for Test-1 (crowd-noise-added speech, SNR: 10dB).

また Test-2 に対し適応実験を行った。雑音クラスタの数が 28 で各音素クラスタ数を変えた時、平均だけを適応する場合と、平均と分散両方を適応する場合の結果を図 12 に示す。1 クラスタで平均と分散の両方を適応する場合に最も高い精度が得られた。最後に 3 種類の雑音クラスタの条件で平均と分散両方を適応する実験を行った。結果を図 13 に示す。16 クラスタの場合に最も高い精度が得られた。

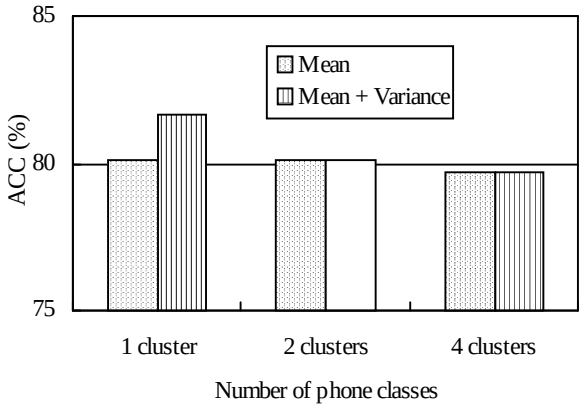

Fig. 9: Recognition results using PLT for Test-1 (crowd-noise-added speech, SNR: 15dB).

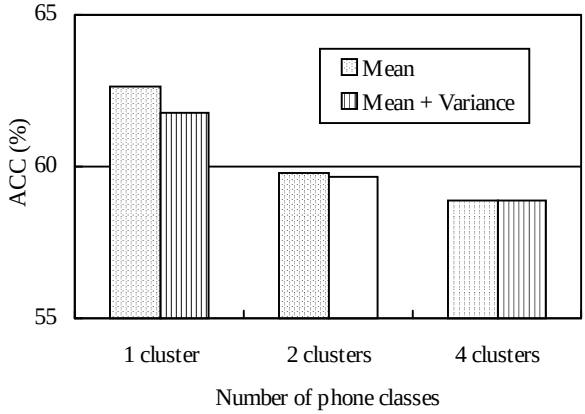

Fig. 10: Recognition results using PLT for Test-1 (exhibition hall noise-added speech, SNR: 10dB).

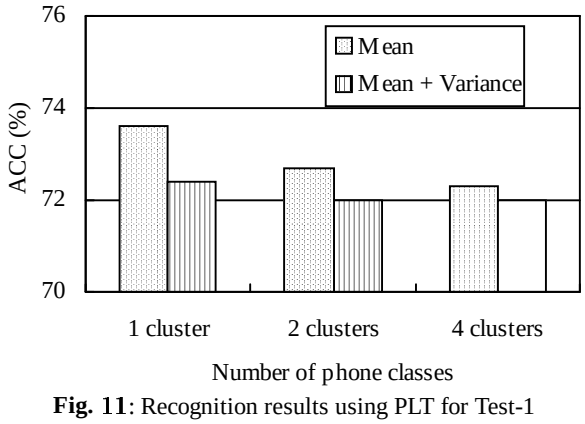

Fig. 11: Recognition results using PLT for Test-1 (exhibition hall noise-added speech, SNR: 15dB).

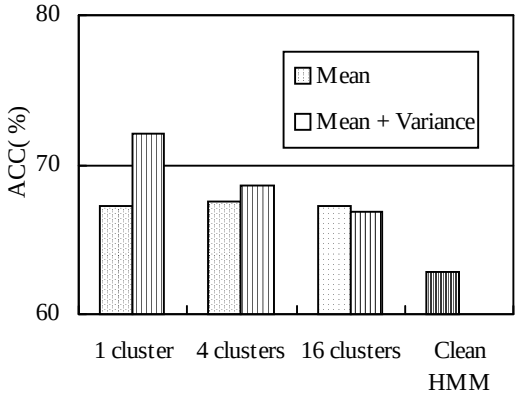

Fig. 12: Recognition results using PLT for Test-2 (number of noise clusters: 28).

以上の結果から、人工的に単一雑音を付加した音声(Test-1)より,音声に多種類の雑音が加わり、雑音が変動しているような実環境の場合(Test-2)に分散適応の効果が大きいことが確認された。ベースラインに比べ、単語正解精度は約 10%増加した。なおこの結果は、以前の HMM 合成法やニューラルネットワークを用いた手法による結果よりも、認識率の向上において優れていることが確認されている。

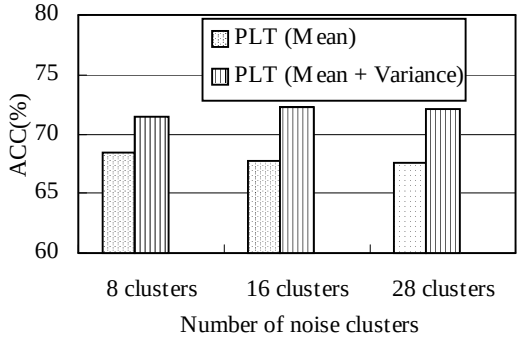

Fig. 13: Recognition results using PLT for Test-2 (number of phone clusters: 1).

5. まとめ

音声認識における耐性向上を目指して、音響モデルすなわち音素 HMM の雑音適応の新しい方法を提案した。

雑音と SNR の組合せについて、クラスタ選択を行ってから、各クラスタごとに線形変換を適用する。クラスタの選択には計算量を考慮し、GMM を用いた。評価実験によって提案手法の有効性を確認した。線形変換に関しては、音素数を 1 クラスタにして、平均と分散両方を適応する方法が最も有効であることが分かった。

音声認識における耐性向上は極めて重要な研究課題であり、音声認識が実際の場で広く実用化されるためには、ハンズフリー入力を前提とした研究などの一層の推進が必要である。

参考文献

[1] F. Martin, et al., “Recognition of noisy speech by composition of hidden Markov models”, *Proc. Eurospeech*, pp. 1031-1034 (1993).

[2] S. Furui and D. Itoh, “Noise adaptation of HMMs using neural networks”, *Proc. ISCA ITRW ASR2000*, pp. 160-167 (2000).

[3] Y. Minami and S. Furui, “A maximum likelihood procedure for a universal adaptation method based on HMM composition”, *Proc. ICASSP*, pp. 129-132 (1995).

[4] C. Lawrence, et al., “Integrated bias removal techniques for robust speech recognition”, *Computer Speech and Language*, vol.13, No.3, pp. 283-298. (1999).

[5] Y. Gong, et al., “Transforming HMMs for speaker-independent hands-free speech recognition in the car”, *Proc. ICASSP*, vol.1, pp. 297-300 (1999).

[6] 管村 他, “SPLIT マルチテンプレート法による不特定話者単語音声認識”, 信学技報, S82-64 (1982).

[7] M.J.F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework”, *Computer Speech and Language*, vol.10, No.3, pp. 249-264 (1996).

[8] 桜井 他, “ニュース音声認識における言語モデルの改良”, 音学春季講論, pp.57-58 (1999).