Vol.2011-SLP-89 No.22
2011/12/20

高即応・高精度な歪み特徴量モデルの推定のための動的静的アプローチ

吉岡拓也[†1]　中谷智広[†1]

本稿では，特徴量空間における歪み過程のモデルを，雑音の変化に感度よく，かつ精度よく推定するための動的静的アプローチを提案する．歪みによって劣化した特徴量からクリーン特徴量を直接推定する特徴量強調は，環境変動に頑健な音声認識に向けた有望なアプローチである．しかしながら，従来の方法では，歪み過程のモデルとして定常ないし変化の遅い加法性雑音を仮定しており，変化の速い雑音に対する効果は限定的であった．動的静的アプローチは，この問題を克服するために，動的ステップと静的ステップの 2 つのステップで構成される．動的ステップでは，各時間フレームにおける雑音の特徴量を一次的に推定する．動的ステップの目的は雑音の変化の特性を捉えることであり，そのために推定は波形や高次元スペクトルの空間で行われる．静的ステップでは，動的ステップで得られた雑音特徴量の一次推定値をクリーン特徴量モデルを用いて補正すると同時に，乗法性雑音の特徴量も最尤法により推定する．本稿では，マイクロホンアレイによる雑音下音声認識と残響音声認識の二つのシナリオについて動的ステップの構成方法を示し，実験により有効性を示す．

Dynamic-Static Approach to Estimation of Change-Sensitive and Accurate Feature Corruption Model

Takuya Yoshioka[†1] and Tomohiro Nakatani[†1]

This paper proposes a novel approach to the estimation of a change-sensitive and accurate feature corruption model, called the dynamic-static approach. Feature enhancement estimates clean features based only on features corrupted by acoustic interferences and is known to effectively improve the speech recognition performance in acoustically adverse environments. However, existing feature enhancement methods use a feature corruption model assuming stationary or slowly varying noise, which precludes its use in highly non-stationary noise environments. The dynamic-static approach overcomes this problem by using two steps called a dynamic (D) step and a static (S) step. The D step roughly estimates a noise waveform or spectrum at each frame and transforms it to the feature space. The aim of the D step is to capture the dynamics of noise change, which is why the estimation is performed in the waveform or spectrum space. The S step compensates for the estimation errors of the noise features and estimates multiplicative distortion parameters simultaneously by using a clean feature model. Experimental results show that the dynamic-static approach significantly reduces word error rates in both microphone array-based noisy speech recognition and single-channel reverberant speech recognition tasks.

1. はじめに

音声認識技術の実用化の勢いが目覚ましい．音声認識を用いたアプリケーションが数多く開発されるようになってきた．こうした普及の背景には，技術の継続的な洗練化，計算機能力の増大，情報端末の小型化・複雑化に伴う音声インタフェースに対する需要の高まり，マルチメディアデータの爆発的な増大などがある．こうした流れに伴って，広範な分野において音声認識に対する需要と期待される技術水準が高まっており，更なる技術改良が求められている．特に，現在の多くの音声認識アプリケーションは近接マイクロホンの利用を仮定しているが，遠隔発話環境でも音声認識に対する需要は高い．実際，会議音声認識，コンシューマ生成ビデオへの自動アノテーション，テレビ会議における自動音声翻訳，ハンズフリー音声インタフェース等，遠隔発話音声認識によってはじめて実現されるアプリケーションは多数ある．すなわち，遠隔発話音声認識技術は，音声認識の応用範囲を拡大する上で極めて重要である．

遠隔発話環境では，他音源による干渉と残響の両方の歪みによって認識性能が著しく低下する．したがって，これらの歪み要因に対して音声認識システムを頑健にすることが求められる．頑健化へのアプローチにはいくつかあるが（解説記事としては 1) が挙げられる），本研究では特徴量強調に着目する．特徴量強調では，観測された劣化音声の特徴量からクリーン音声の特徴量を推定する．特徴量強調の代表的な方法として，VTS（vector Taylor series）強調がある．典型的な VTS 強調の方法では，混合正規分布（GMM: Gaussian mixture model）等の予め学習されたクリーン音声の特徴量のモデルと，観測データから推定された歪み過程のモデルを用いて，劣化特徴量が与えられたときのクリーン特徴量の事後分布を求め，その平均をクリーン特徴量の推定値とする．この方法は，比較的小さい計算量で効果的に認識性能を向上できる[1)]．

しかしながら，従来の特徴量強調方法では，速く変化する歪みに対して十分な歪み補正効

†1 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

果が得られない．これは，従来の方法が，歪みの要因として定常ないし変化の遅い加法性雑音を仮定するためである[2)–4)]．ところが現実には，遠隔発話環境で生じる雑音は，他話者の音声や残響のように時間フレーム単位で大きく変化する（本研究では，残響—厳密には後部残響—を非定常な加法性雑音として捉える）．これが，遠隔発話音声認識の精度が未だ実用的な水準に遠く及ばない要因の一つである．

こうした現状を克服するため，本稿では，時間フレーム毎の歪み過程モデルのパラメータ（具体的には加法性雑音の特徴量の正規分布パラメータと乗法性雑音の特徴量）を精度よく推定するための動的静的アプローチを提案する．提案方法は，二つの考え方を背景としている．一つは，特徴量空間よりも波形ないしスペクトル（本稿では，対数パワースペクトルの意味でスペクトルという語を使う）の空間を用いて雑音を推定したほうが，雑音の変化を捉えやすいことである．これは，波形やスペクトル空間のほうが，次元がずっと高く，室内の音の伝播特性を用いられるため，信号の詳細な構造を捉えるのに適しているからである．もう一つは，特徴量空間で歪み過程モデルを推定したほうが，クリーン特徴量モデルと親和性の高いモデルを得られることである．これは，特徴量空間では，歪み過程モデルのパラメータ推定にクリーン特徴量モデルを直接用いられるからである．

そこで，動的静的アプローチでは，動的ステップと静的ステップからなる二つのステップによって時間フレーム毎の歪み過程モデルを推定する．動的ステップの目的は，加法性雑音の変化特性を捉えることである．具体的には，動的ステップでは，波形ないしスペクトルの空間で加法性雑音系列の一次的な推定値を求めて，特徴量に変換する．一方，静的ステップの目的は，加法性雑音特徴量系列の真値と動的ステップで求められた一次推定値の間の誤差をクリーン特徴量モデルを用いて補正し，合わせて誤差の分散と乗法性雑音の特徴量も推定することである．これにより，時間フレーム毎の歪み過程モデルのパラメータが得られる．具体的には，各時間フレームにおける加法性雑音の真値の一次推定値からの誤差は正規分布からの独立な標本であると仮定し，その平均と分散，及び乗法性雑音をクリーン特徴量モデルを用いて最尤推定する（本稿では，特徴量とスペクトルのいずれに言及しているかが文脈から明らかな場合，単に「加法性雑音」のように「特徴量」や「スペクトル」などと明示しない）．各時間フレームにおける加法性雑音の正規分布は，この誤差正規分布の平均を当該フレームにおける加法性雑音でシフトしたものとして得られる．その後，求められた歪み過程モデル（加法性雑音の正規分布と乗法性雑音）を用いて，時間フレーム毎に VTS 強調を行うことでクリーン特徴量を推定する．提案方法の重要な点は，雑音の変化特性の推定は動的ステップで行い，静的ステップでは時不変なパラメータのみ推定するということである．これにより，静的ステップにおける精度よいパラメータ推定を可能にしている．

動的静的アプローチの利点は，汎用性の高さである．加法性雑音を一次的に推定する動的ステップには，目的や環境に応じて最適な方法を選ぶことができる．具体例として，4 章では，遠隔発話環境における二つの異なるシナリオを想定した動的ステップの構成方法を示す．一つめは複数話者が同時に発話している環境での音声認識であり，マイクロホンアレイを用いて動的ステップを構成する．二つめは残響音声の認識であり，既存の後部残響推定方法を用いて動的ステップを構成する．5 章では，両方の事例について評価実験を行い，提案方法の有効性を示す．

本稿では以下，2 章で VTS 強調の方法を概説した後，3 章で提案方法における静的ステップのアルゴリズムについて説明する．また，4 章では，提案方法の適用シナリオを二つ検討し，それぞれにおける動的ステップの構成方法を与える．5 章で評価実験について述べ，6 章で本稿で得られた知見をまとめる．

2. VTS 強調のあらまし

本稿では簡単のため，特徴量として対数メル周波数スペクトルを用いて提案方法を説明する．劣化した音声波形とクリーンな音声波形をそれぞれ $x(t)$，$s(t)$ と書く．また，これらから抽出される特徴量を，それぞれ $x_{n,j}$，$s_{n,j}$ と書く．t，n，j はそれぞれ，時間領域での標本のインデクス，時間フレームのインデクス，メル周波数チャネルのインデクスである．また，発話区間を $\{t; 1 \le t \le T\}$，処理対象の周波数帯域を $\{j; 1 \le j \le F\}$ と表す．

特徴量強調の目的は，クリーン特徴量 $s_{n,j}$ の最小平均二乗誤差（MMSE: minimum mean squared error）推定値を求めることである．MMSE 推定値は，劣化特徴量 $x_{n,j}$ が与えられたときの $s_{n,j}$ の事後分布の平均 $E(s_{n,j}|x_{n,j})$ として求められる．ただし，E は期待値演算を表す．VTS 強調では，以下で述べるクリーン特徴量と劣化過程の各モデルを用いてこの事後分布を求める．

クリーン特徴量モデルには通常，対角共分散行列をもつ GMM が用いられる．すなわち，$\boldsymbol{s}_n = [s_{n,1}, \cdots, s_{n,F}]$ について，次式の確率密度関数を仮定する．

$$p(\boldsymbol{s}_n) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{F} p(s_{n,j}|k) \tag{1}$$

$$p(s_{n,j}|k) = f_{\mathrm{N}}(s_{n,j}; \nu_{k,j}, \tau_{k,j}^2) \tag{2}$$

K は GMM の混合数，$f_{\mathrm{N}}(x; \mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布の確率密度関数を表す．この GMM のパラメータ $\{\pi_k, \nu_{k,j}, \tau_{k,j}^2\}_{1 \le k \le K, 1 \le j \le F}$ は，クリーン音声のコーパスから予め学習さ

れる．

劣化過程モデルは，クリーン特徴量 $s_{n,j}$ から劣化特徴量 $x_{n,j}$ が生成される過程をモデル化したものである．このモデルはミスマッチ関数，加法性雑音の特徴量のモデル，及び乗法性雑音の特徴量で構成される．ミスマッチ関数は，劣化特徴量 $x_{n,j}$，クリーン特徴量 $s_{n,j}$，加法性雑音特徴量 $r_{n,j}$，乗法性雑音特徴量 $h_{n,j}$ の関係を表現し，具体的には次式の非線形関数 f で定義される（インデクス n と j は見やすさのため省略する）．

$$x = f(s, r, h) = s + h + \log(1 + \exp(r - s - h)) \tag{3}$$

本稿では，乗法性雑音 $h_{n,j}$ は発話中には変化しないと仮定し，以後 h_j と表記する．加法性雑音特徴量 $r_{n,j}$ は，平均 $\mu_{n,j}$，分散 $\sigma^2_{n,j}$ の正規分布を用いてモデル化される．

$$p(r_{n,j}) = f_{\mathrm{N}}(r_{n,j}; \mu_{n,j}, \sigma^2_{n,j}) \tag{4}$$

本稿では，分散 $\sigma^2_{n,j}$ については時不変なパラメータとして扱うこととし，以後 σ^2_j と表記する．一方，平均 $\mu_{n,j}$ は時間フレーム n に依存して変化してもよい．劣化過程モデルはパラメータ $\{\mu_{n,j}, \sigma^2_j, h_j\}_{1\le n\le N, 1\le j\le F}$ の値に依存して決まる．これらの値は劣化音声から直接推定される．

提案方法と従来方法との違いは $\mu_{n,j}$ の扱い方にある．従来の特徴量強調方法では $\mu_{n,j}$ も時不変ないしゆっくりと変化すると仮定していたため，早く変化する歪みに対する補正効果が不十分であった．これに対して，次章で述べるように，提案方法では動的ステップで $\mu_{n,j}$ の変化特性を推定することで，そうした制約を回避する．

以上のモデルの下で，クリーン特徴量の MMSE 推定値 $\hat{s}_{n,j}$ は次式によって求めることができる．

$$\hat{s}_{n,j} = \sum_{k=1}^{K} E(s_{n,j}|x_{n,j}, k) p(k|\boldsymbol{x}_n) \tag{5}$$

ミスマッチ関数 f の非線形性により，(5) の右辺は解析的に求まらない．そこで，VTS 強調では，非線形関数 f をクリーン特徴量モデルの平均 ν，加法性雑音特徴量モデルの平均 μ，及び乗法性雑音 h のまわりで線形化する．これによって，右辺の二つの項の近似値を解析的に求める．詳しいアルゴリズムについては，1) の 33.6.2 節を参照されたい．

3. 動的静的アプローチ

提案する動的静的アプローチによる特徴量強調処理の流れを**図 1** に示す．まず，動的ステップにおいて，マイクロホンで観測された劣化音声から時間フレーム毎の加法性雑音のスペクトルを一次的に推定する（これを参照スペクトルと呼ぶ）．同時に，各時間フレームにおける観測スペクトルも求める．これらのスペクトルは，メルフィルタバンクを適用する前の高次元データである．次に，特徴抽出ステップにおいて，観測スペクトルと参照スペクトルから特徴量を抽出し，観測特徴量と参照特徴量（加法性雑音特徴量の一次推定値に相当）の各時系列を求める．静的ステップでは，これらの特徴量系列とクリーン特徴量モデルを用いて，加法性雑音特徴量モデルの時間フレーム毎の平均 $\{\mu_{n,j}\}_{1\le n\le T, 1\le j\le F}$，分散 $\{\sigma^2_j\}_{1\le j\le F}$，及び乗法性雑音の特徴量 $\{h_j\}_{1\le n\le F}$ を推定する．VTS 強調ステップでは，これら歪み過程モデルのパラメータの推定値とクリーン特徴量モデルを用いて，時間フレーム毎に強調特徴量を求める．

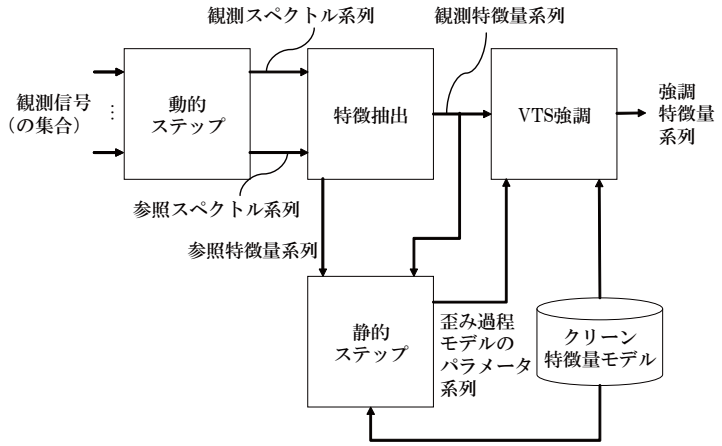

図 1 動的静的アプローチによる特徴量強調処理の流れ

動的ステップの実現方法は，目的や環境に応じて選択する．例えばマイクロホンアレイが利用可能な場合，独立成分分析[5]（ICA: independent component analysis）や時間周波数マスキング[6] 等の方法を用いて参照スペクトル系列を求めることができる．あるいは，残響によって劣化した特徴量から残響の影響を取り除く場合には，既存の後部残響推定方法を用いることができる[7),8)]．動的ステップについては，4 章で具体的な実現例を延べる．本章では以下，静的ステップについて詳しく説明する．

3.1 静的ステップの定式化

静的ステップの目的は，加法性雑音モデルの時間フレーム毎の平均 $\{\mu_{n,j}\}_{1\le n\le T, 1\le j\le F}$，分散 $\{\sigma^2_j\}_{1\le j\le F}$，及び乗法性雑音 $\{h_j\}_{1\le n\le F}$ を推定することである．一般に，観測特徴量系列

$\{x_{n,j}\}_{1\le n\le T,1\le j\le F}$ しか与えられない場合，未知パラメータが観測データよりも多くなるのでこの問題は解けない．これが，従来の特徴量強調方法において，平均 $\mu_{n,j}$ が時間に依存しない，あるいはゆっくりとしか変化しないと仮定せざるを得なかった理由である．しかしながら，この仮定は，非定常性の大きい歪みに対する補正効果を限定的にしてしまう．提案方法では，加法性雑音特徴量の一次的な推定値である参照特徴量系列 $\{\hat{r}_{n,j}\}_{1\le n\le T,1\le j\le F}$ も用いることで，この問題を解決する．

提案する動的静的アプローチの要諦は，加法性雑音モデルの平均 $\mu_{n,j}$ を参照特徴量 $\hat{r}_{n,j}$ と時不変なバイアス b_j に分解する点である．

$$\mu_{n,j} = \hat{r}_{n,j} + b_j \tag{6}$$

この仮定によって，静的ステップでは加法性雑音モデルの平均バイアス $\{b_j\}_{1\le j\le F}$，分散 $\{\sigma_j^2\}_{1\le j\le F}$，及び乗法性雑音 $\{h_j\}_{1\le n\le F}$ のみを推定すればよい．この問題は，観測データよりも未知パラメータ数が少ないので解ける．

平均バイアス b_j は加法性雑音の真値 $r_{n,j}$ の一次推定値 $\hat{r}_{n,j}$ からの平均的な誤差と解釈できる．このとき，σ_j^2 は誤差の分散に相当する．すなわち，b_j と σ_j^2 を特徴量空間で推定することは，クリーン特徴量を使って $\hat{r}_{n,j}$ の誤差を補正していると見做せる．

本稿では，未知パラメータ $\Theta = \{b_j, \sigma_j^2, h_j\}_{1\le j\le F}$ を最尤法によって推定する．すなわち，$\mathbb{X} = \{x_{n,j}\}_{1\le n\le T,1\le j\le F}$ とすると，尤度関数 $p(\mathbb{X};\Theta)$ を最大化する $\hat{\Theta}$ を求める．この尤度関数は次式のように分解される．

$$p(\mathbb{X};\Theta) = \prod_{n=1}^{T}\prod_{j=1}^{F}\sum_{k=1}^{K}\pi_k p(x_{n,j}|k;\Theta) \tag{7}$$

右辺の確率密度関数 $p(x_{n,j}|k;\Theta)$ は，VTS 強調と同様にミスマッチ関数 f を線形化して正規分布近似したものを用いる．

$$p(x_{n,j}|k,;\Theta) = f_{\mathrm{N}}(x_{n,j};\psi_{n,k,j},\upsilon_{n,k,j}^2) \tag{8}$$

この正規分布の平均 $\psi_{n,k,j}$ と分散 $\upsilon_{n,k,j}^2$ は次式で与えられる．

$$\psi_{n,k,j} = f(\nu_{k,j},\hat{r}_{n,j}+b_j,h_j) \tag{9}$$

$$\upsilon_{n,k,j}^2 = g(\nu_{k,j},\hat{r}_{n,j}+b_j,h_j)^2\tau_{k,j}^2 + (1-g(\nu_{k,j},\hat{r}_{n,j}+b_j,h_j))^2\sigma_j^2 \tag{10}$$

$g(s,r,h)$ は非線形関数 $f(s,r,h)$ の s に関する偏導関数であり，具体的には次式のように書ける．

$$g(s,r,h) = \frac{1}{1+\exp(r-s-h)} \tag{11}$$

以上で，静的ステップで解くべき問題が定式化された．

3.2 EM アルゴリズムによる最適化

本稿では，加法性雑音モデルの平均バイアスと分散，及び乗法性雑音の最尤解を求めるために，二重 EM アルゴリズム[9),10)] の変形を用いる．提案アルゴリズムでは，加法性雑音モデルのパラメータ（平均バイアスと分散）の推定と乗法性雑音の推定を交互に繰り返す．本稿では，平均バイアスと分散を推定するアルゴリズムについてのみ示す．乗法性雑音の推定アルゴリズムも同様にして導ける．

提案アルゴリズムでは，各時間フレームにおける GMM の要素分布を示すインデクスと加法性雑音を潜在変数と見做す．すなわち，$\hat{\Theta}$ を Θ の現在の推定値とすると，EM アルゴリズムの各ループで最大化される補助関数は次式で与えられる．

$$Q(\Theta;\hat{\Theta}) = \sum_{n=1}^{T}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})\sum_{j=1}^{F}\int q_{n,k,j}(r;\hat{\Theta})\log f_{\mathrm{N}}(r;\hat{r}_{n,j}+b_j,\sigma_j^2)dr \tag{12}$$

$\gamma_{n,k}(\hat{\Theta})$ は，現在のパラメータ推定値の下で，時間フレーム n において k 番目の要素分布が選択される事後確率である．$q_{n,k,j}(r;\hat{\Theta})$ は，現在のパラメータ推定値と要素分布のインデクス k が与えられた下での，時間フレーム n，メル周波数チャネル j における加法性雑音の事後確率密度関数である．$\gamma_{n,k}(\hat{\Theta})$ と $q_{n,k,j}(r;\hat{\Theta})$ は E ステップにおいて次式にしたがって計算される．

$$\gamma_{n,k}(\hat{\Theta}) = \frac{\prod_{j=1}^{F}p(x_{n,j}|k;\hat{\Theta})}{\sum_{k=1}^{K}\prod_{j=1}^{F}p(x_{n,j}|k;\hat{\Theta})} \tag{13}$$

$$q_{n,k,j}(r;\hat{\Theta}) = f_{\mathrm{N}}(r;\kappa_{n,k,j}(\hat{\Theta}),\lambda_{n,k,j}^2(\hat{\Theta})) \tag{14}$$

$$\kappa_{n,k,j}(\hat{\Theta}) = \hat{r}_{n,j} + \hat{b}_j + \frac{\hat{g}_{n,k,j}\hat{\sigma}_j^2(x_{n,j}-\hat{\psi}_{n,k,j})}{\hat{\upsilon}_{n,k,j}^2} \tag{15}$$

$$\lambda_{n,k,j}^2(\hat{\Theta}) = \frac{\hat{\sigma}_j^2\hat{g}_{n,k,j}^2\tau_{k,j}^2}{\hat{\upsilon}_{n,k,j}} \tag{16}$$

$$\hat{g}_{n,k,j} = g(\nu_{k,j},\hat{r}_{n,j}+\hat{b}_j,\hat{h}_j) \tag{17}$$

M ステップでは，これらを用いて加法性雑音特徴量モデルの平均バイアスと分散の推定値を更新する．

$$\hat{b}_j = \frac{1}{T}\sum_{n=1}^{T}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})(\kappa_{n,k,j}(\hat{\Theta})-\hat{r}_{n,j}) \tag{18}$$

$$\hat{\sigma}_j^2 = \frac{1}{T}\sum_{n=1}^{T}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})(\kappa_{n,k,j}(\hat{\Theta})^2+\lambda_{n,j,k}^2(\hat{\Theta})) - \hat{b}_j^2 \tag{19}$$

これら E ステップと M ステップを繰り返すことで，歪み過程モデルの最終的なパラメータ推定値を得る．なお，EM アルゴリズムのループを一巡する度に f の線形化の中心が変化す

Vol.2011-SLP-89 No.22
2011/12/20

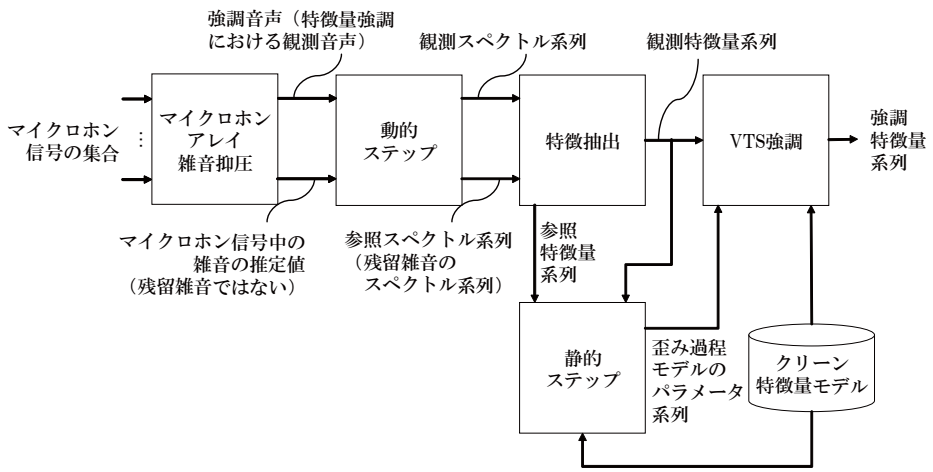

図 2 マイクロホンアレイを用いた雑音下音声認識への動的静的アプローチの適用

るため，このアルゴリズムは必ずしも尤度関数の単調増加を保証しない．単調増加を保証するために，バックオフ[11] を用いることもできる．現在の著者らの実装では，バックオフを用いている．バックオフについては 11) を参照されたい．

4. 動的ステップの構成例

本章では，動的静的アプローチの二つの適用シナリオを検討し，それぞれにおける動的ステップの構成方法を具体的に述べる．4.1 節では，マイクロホンアレイを用いた雑音下音声認識について検討する．4.2 節では，残響音声認識を取り上げる．

4.1 マイクロホンアレイを用いた雑音下音声認識

このシナリオでは，**図 2** に示すように，マイクロホンアレイによる雑音抑圧の後処理として動的静的アプローチによる特徴量強調を行う．収音装置としてマイクロホンアレイを用いる場合，音声認識に先立って ICA や時間周波数マスキング等の雑音抑圧技術によって目的話者の音声だけを予め強調しておくことができる．これによって音声認識システムに入力される音声の信号対雑音比（SNR: signal-to-noise ratio）が改善されるため，認識精度もある程度向上する．特に，マイクロホンアレイに基づく雑音抑圧は，単一マイクロホンによる方法と比べて，非定常な雑音に対して高い効果をもつ．しかしながら，残響や話者の移動等の要因により，実際には強調された音声には雑音がある程度残留するため，十分な認識性能が達成されない。したがって，マイクロホンアレイを用いる場合でも，この残留雑音の影響を補正することが必要になる．マイクロホンアレイで観測された音声に含まれる雑音が非定常である場合，残留雑音も非定常になりがちであるため，従来の特徴量強調方法では必ずしも十分な性能改善が得られない（この問題は，ICA 等の線形フィルタを用いる場合に顕著である）．この問題を解決するのに動的静的アプローチを用いる．

提案アルゴリズムでは，マイクロホンアレイ雑音抑圧システムは目的音声と加法性雑音の両方の信号を出力するという仮定を用いて，動的ステップを構成する．ICA や時間周波数マスキングを含む多くの雑音抑圧処理技術はこの仮定を満たす．

具体的には，提案する動的ステップでは，この仮定に基づいてマイクロホンアレイ処理で求められた強調音声に残留する雑音成分を抽出するためのバイナリマスク[12] を推定し，これを強調音声に適用することで残留雑音のスペクトルの一次推定値を求める．バイナリマスクは，音声分離で広く使われているアプローチである．今，i をメルフィルタバンク処理前のスペクトル空間における周波数ビンのインデクスとする．バイナリマスク $A_{n,i}$ は各時間フレーム，各周波数ビンに対して定義される 2 値変数であり，$A_{n,i} = 0$ ならば当該時間周波数点において目的音声が存在し，$A_{n,i} = 1$ ならば目的音声が存在しないことを示す．先ほどの仮定を用いると，$X_{1,n,i} < X_{2,n,i}$ ならば $A_{n,i} = 1$，$X_{1,n,i} \geq X_{2,n,i}$ ならば $A_{n,i} = 0$ とすればよい．ただし，$X_{1,n,i}$ と $X_{2,n,i}$ は，それぞれマイクロホンアレイに基づく強調音声と雑音のスペクトルを表す．こうして計算されたバイナリマスクを用いて，残留雑音スペクトルの一次推定値，すなわち参照スペクトル $\hat{R}_{n,i}$ を強調音声のスペクトル $X_{n,i}$ の局所重みつき平均として計算する．

$$\hat{R}_{n,i} = \frac{\sum_{\tau=-\Delta_T}^{\Delta_T} \sum_{\phi=-\Delta_F}^{\Delta_F} A_{n+\tau,i+\phi} X_{1,n+\tau,i+\phi}}{\sum_{\tau=-\Delta_T}^{\Delta_T} \sum_{\phi=-\Delta_F}^{\Delta_F} A_{n+\tau,i+\phi}} \tag{20}$$

ただし，Δ_T と Δ_F はそれぞれ局所平均を求めるための窓幅であり，現在の実装では，$\Delta_T = 3$，$\Delta_F = 2$ としている．こうして得られた参照スペクトルに対してメルフィルタバンクを適用することで，参照特徴量 $\hat{r}_{n,j}$ を得る．

4.2 残響音声認識

このシナリオでは，単一のマイクロホンで観測された残響を含む音声を認識する．このために，残響による歪みを特徴量領域で補正する．残響に対する既存の特徴量強調方法としては，13) や 14) がある．前者はケプストラムの空間における畳み込みとして残響をモデル化しているが，このモデルの精度は高くないため，効果は限定的である．後者は精度はよ

いものの，クリーン特徴量と残響の両方のモデルが複雑である．これに対して，本節では，極めて単純でかつ効果的な動的ステップの構成方法を与える．

提案する動的ステップでは，後部残響のスペクトルを Lebart らのモデル[7] を用いて一次的に推定し，これを参照スペクトルとする．具体的には，参照スペクトル $\hat{R}_{n,i}$ は次式にしたがって求められる．

$$\hat{R}_{n,i} = X_{n-\Delta,i} + \alpha \tag{21}$$

Δ は 50〜100 ミリ秒程度に相当するように決められる．現在の実装では，約 65 ミリ秒となるように Δ を設定している．α は部屋の残響時間に依存した定数であるため，従来は (21) を用いる前に残響時間を推定する必要があった．提案方法では，静的ステップの平均バイアスによって α のずれは特徴量領域で自動的に補正されるため，この値は適当に決めてよい．現在の実装では，$\alpha = 0.8$ としている．(21) で求められた参照スペクトルに対してメルフィルタバンクを適用することで，参照特徴量 $\hat{r}_{n,j}$ を得る．

5. 実験結果

本章では，4 章で取り上げた二つのシナリオについて実験を行い，動的静的アプローチの効果を示す．

5.1 マイクロホンアレイによる複数話者混在環境での音声認識

最初の実験は，複数話者が混在する環境での連続数字認識である．具体的には，目的話者が発声した連続数字と別の話者の音声が混合されたステレオ音声から，目的話者の数字だけを認識する．

実験試料となる混合音声は，Aurora2 テストセットに含まれるクリーンな連続数字音声 4004 個と TIMIT テストセットから抽出したクリーン音声を，室内で測定したインパルス応答を用いて混合することで作成した．インパルス応答は，二個のマイクロホンを用いて残響時間が 0.13 秒に設定された可変残響室で測定した．残響室の大きさは，幅 4.45 m，奥行き 3.35 m，高さ 2.5 m である．二個のマイクロホンは残響室の中央に近接して設置し，これから 1m の距離に 2 個のスピーカを配置した．各スピーカはマイクロホンの左側ないし右側 30 度に位置し，それぞれ目的話者と干渉話者に対応する．以上の手続きにより，合計 4004 個のステレオ混合音声を作成した．マイクロホンアレイ雑音抑圧の方法として ICA を用いた．

音響モデルは，Aurora2 で標準的に使われている complex backend のクリーンモデルを用いた．すなわち，音響モデルは 16 状態 3 混合の話者非依存の単語 HMM で構成される．音声認識に用いる特徴量には，C0〜C12 の MFCC とその変化速度，加速度の合計 39 次元を用いた．また，同じ Aurora2 学習データから，特徴量強調に用いる GMM を作成した．GMM の混合数は 1024 とした．

表 1 マイクロホンアレイによる複数話者混在環境での数字誤り率

雑音抑圧なし	ICA	ICA + 従来の VTS 強調	ICA + 提案方法
178.25%	25.78%	20.41%	3.21%

表 2 残響環境における数字誤り率

残響時間	残響補正なし	残響補正あり（提案方法）	残響補正あり（真の後部残響を用いた上限値）
0.6 秒	29.23%	14.11%	8.89 %
0.5 秒	19.04%	7.01%	4.92 %
0.4 秒	13.18%	5.40%	3.94 %
0.3 秒	9.94%	3.96%	3.07 %
0.2 秒	2.71%	2.17%	1.66 %

表 1 に数字誤り率を示す．ICA により数字誤り率は改善されたものの，依然として 25.78%という高い水準であった．これは，目的話者が強調された音声には他話者の音声が依然として残留していた影響である．ICA で強調された音声に対して従来の VTS 強調方法を用いた場合，数字誤り率は 20.41%であり，その効果は限定的であった．これは，残留雑音が非定常であることによると考えられる．一方，提案した動的静的アプローチを用いて特徴量強調した場合，3.21%の数字誤り率を達成した．この結果は，提案方法が他話者の残留音声という極めて非定常な雑音による歪みを効果的に補正できることを示している．

5.2 残響音声認識

次に，モノラル残響音声の数字認識実験を行った．実験試料となる残響音声は，先の実験と同様，Aurora2 テストセットに含まれるクリーンな連続数字音声に室内で測定されたインパルス応答を畳み込んで作成した．インパルス応答は先の実験と同じく，可変残響室内で測定した．この実験では，残響時間が 0.2〜0.6 秒になるように壁面を調整し，スピーカとマイクロホンの距離は 2.5 m とした．

表 2 にこの実験における数字誤り率を表示する．いずれの残響時間においても，提案方法を用いて残響の影響を補正することで数字誤り率が大きく低減された．例えば残響時間が 0.6 秒の場合，残響音声を直接認識した場合の数字誤り率は 29.23%であったのに対して，提案方法により 14.11%まで数字誤り率を低減できた．

しかしながら，認識誤りを大幅に削減できたものの，未だ十分な性能には達していない．この原因について調べるために，参照スペクトル $\hat{R}_{n,l}$ として (21) の代わりに真の後部残響を

用いた場合の数字誤り率を求めた．ただし，真の後部残響はインパルス応答のうち直接音の到達から 50 ミリ秒以降の部分をクリーン音声に畳み込むことで求めた．結果を表 2 の最右欄に示す．例えば残響時間が 0.6 秒の場合，この条件での数字誤り率は 8.89%であった．この誤り率は動的静的アプローチが達成しうる限界であり，さらに誤りを削減するためには，他のアプローチと併用する必要があることを示唆している．一方，14.11%と 8.89%のギャップは，動的ステップの構成方法についても改良の余地が残されていることを示している．

6. おわりに

本稿では，雑音の変化に感度よく，かつ精度よく歪み過程モデルのパラメータを推定するための動的静的アプローチについて述べた．この方法の要諦は，波形やスペクトル等の高次元空間で雑音の時系列を一次的に推定し（動的ステップ），これをクリーン特徴量モデルを用いて補正することにある（静的ステップ）．動的ステップでは，信号の詳細な構造や室内での音の伝播特性を利用できるため，雑音の変化の特性を比較的容易に捉えられる．雑音の変化の特性を動的ステップで推定しておくことで，静的ステップでは時不変なパラメータのみ推定すればよい．動的ステップの構成方法は，目的や環境に応じて選択する必要がある．本稿では，マイクロホンアレイに基づく雑音下音声認識と単一マイクロホンによる残響音声認識の二つのシナリオを想定して，動的ステップの構成方法を示した．また，両方のシナリオについて評価実験を行った．いずれにおいても提案方法は認識誤りを効果的に削減できることが確認され，提案方法の幅広い有用性が示唆された．

今後は提案方法の適用範囲をさらに広げていく予定である．具体的には，現在，会話音声認識[15] において提案方法を評価しており，既にその効果を確認している．この結果については別途報告予定である．

謝辞 実装について協力頂いた NTT コミュニケーション科学基礎研究所の元実習生 Emmanuel Y. J. Ternon 氏に感謝する．

参考文献

1) Droppo, J. and Acero, A.: Environmental robustness, *Springer Handbook of Speech Processing* (Benesty, J., Sondhi, M.M. and Huang, Y., eds.), Springer, pp.653–679 (2008).
2) Deng, L., Droppo, J. and Acero, A.: Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition, *IEEE Trans. Speech, Audio Process.*, Vol.11, No.6, pp.568–580 (2003).
3) Stouten, V., Van hamme, H. and Wambacq, P.: Model-based feature enhancement with uncertainty decoding for noise robust ASR, *Speech Comm.*, Vol. 48, No. 11, pp. 1502–1514 (2006).
4) Fujimoto, M. and Nakamura, S.: Sequential non-stationary noise tracking using particle filtering with switching dynamical system, *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, pp.769–772 (2006).
5) Makino, S., Lee, T.W. and Sawada, H.: *Blind speech separation*, Springer (2007).
6) Sawada, H., Araki, S. and Makino, S.: A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures, *Proc. IEEE Worksh. Appl. Signal Process. Audio, Acoust.*, pp.139–142 (2007).
7) Lebart, K., Boucher, J.M. and Denbigh, P.N.: A new method based on spectral subtraction for speech dereverberation, *Acta Acustica united with Acustica*, Vol.87, pp.359–366 (2001).
8) Yoshioka, T., Nakatani, T., Kinoshita, K. and Miyoshi, M.: Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model, *Speech Processing in Modern Communication: Challenges and Perspectives* (Cohen, I., Benesty, J. and Gannot, S., eds.), Springer, pp.151–182 (2010).
9) Zhao, Y. and Juang, B.-H.: A comparative study of noise estimation algorithms for VTS-based robust speech recognition, *Proc. Interspeech*, pp.2090–2093 (2010).
10) Gales, M. J.F.: Model-based approaches to handling uncertainty, *Robust Speech Recognition of Uncertain or Missing Data* (Kolossa, D. and Haeb-Umbach, R., eds.), Springer, pp. 101–125 (2011).
11) Liao, H.: Uncertainty decoding for noise robust speech recognition, PhD Thesis, The University of Cambridge (2007).
12) Yilmaz, O. and Rickard, S.: Blind separation of speech mixtures via time-frequency masking, *IEEE Trans. Signal Process.*, Vol.52, No.7, pp.1830–1847 (2004).
13) Kumar, K. and Stern, R.: Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation, *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, pp. 4282–4285 (2010).
14) Krueger, A. and Haeb-Umbach, R.: Model-based feature enhancement for reverberant speech recognition, *IEEE Trans. Audio, Speech, Lang. Process.*, Vol.18, No.7, pp.1692–1707 (2010).
15) Hori, T., Araki, S., Yoshioka, T., Fujimoto, M., Watanabe, S., Oba, T., Ogawa, A., Otsuka, K., Mikami, D., Kinoshita, K., Nakatani, T., Nakamura, A. and Yamato, J.: Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera, *IEEE Trans. Audio, Speech, Language Process.* (2011). to appear.