残響音声のための音響モデル変換*

◎ 槻武也　西本卓也　　嵯峨山茂樹　（東大・情報理工）

1　はじめに

実環境における音声認識では、周囲雑音や残響などの要因によって認識性能が劣化する。周囲雑音などの加法性雑音に対しては、スペクトル減算法などの雑音抑制手法やモデル合成法 [1] が用いられる。また、回線特性などによる乗法性歪みに対しては CMN 法などによって対処できる。しかし、フレーム長に比べて長い残響による歪みは、過去のフレームの音声信号が残響信号として現在のフレームに重畳するため、同様の対処はできない。

残響による観測信号の歪みに対しては、残響に頑健な音響特徴量を用いる方法 [2] などが提案されている。モデル適応法としては、インパルス応答を与えることで短時間の残響に対応したモデル適応を行なう方法 [3] などが提案されているが、残響時間が分析フレームよりも短いと仮定しており、過去のフレームからの残響は考慮されていない。

本稿では、残響に頑健な音声認識を実現するため、残響特性が与えられたときに、クリーン音声のモデルから残響環境に適応した音響モデルへ変換する手法について報告する。HMM は音声生成の確率モデルであることを利用する。

2　残響音声の定式化

2.1　提案手法の枠組

本研究では、特徴量や音響モデル構造や音声認識アルゴリズムには変更を加えずに、音響モデルのパラメータのみ残響環境に適応させる。クリーン音声の場合に比べて同じ音声認識システムが使え計算量の増加はなく扱いやすい。そのために、過去のフレームからの残響を考慮したモデル適応を行なう。本研究で対象とする音響モデルは、状態間の遷移確率と、各状態ごとの MFCC と ΔMFCC の出力確率分布によって構成される一般的な音素 HMM である。出力確率分布は平均と分散で与えられる正規分布である。また、残響特性はパワーエンベロープ伝達関数として与えられる。

2.2　スペクトル領域における残響表現

実環境における残響の影響は、波形領域では環境のインパルス応答の畳み込みで記述できる。インパルス応答が分析フレームに比べて長い場合、短時間スペクトル領域では各フレームに対して、そのフレームとそれ以前のフレームのスペクトルが伝達特性による歪みを受けて重畳する。これは、次式のようにフレーム単位の畳み込みによって近似できる [4]。

$$O(\omega,n) = H(\omega,n) * S(\omega,n) \quad (1)$$

$$= \sum_{l\geq 0} H(\omega,l)S(\omega,n-l) \quad (2)$$

*“Acoustic Model Conversion for Reverberant Speech Recognition” by Takeya SAIKACHI, Takuya NISHIMOTO, and Shigeki SAGAYAMA (The University of Tokyo).

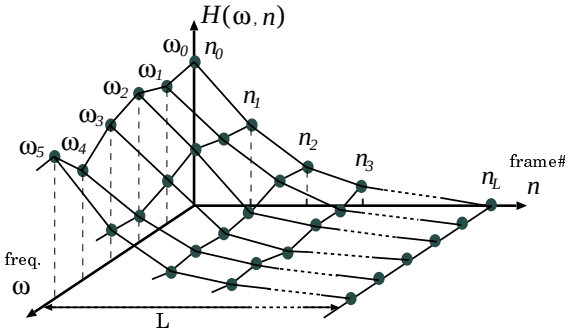

図 1: 残響特性の短時間周波数応答による表現

ここで、n はフレーム番号、ω は周波数（フィルタバンク分析の場合はフィルタバンク番号に対応）、$O(n,\omega)$、$S(n,\omega)$ はそれぞれ観測信号、音声信号の短時間スペクトルである。$H(n,\omega)$ はインパルス応答の短時間スペクトルで、図 1 に模式的に示すようにフレームおよび周波数帯域ごとのパワーの減衰を表す。式 (2) の右辺で加算される各項の $H(\omega,l)S(\omega,n-l)$ は、フレーム n からみて l フレーム前に発声した音声に起因する残響成分に相当する。

3　残響音声のためのモデル変換

3.1　先行音素の情報が既知の場合のモデル変換

まず、過去のフレームの音素と状態継続時間が既知である場合 (具体的には、2 パス探索の第 2 パスの尤度再計算の場合など) に、各フレームの MFCC 分布を予測する方法を導く。

第 n フレームから l フレーム遡った過去の音声の残響成分は、式 (2) の右辺中の $H(\omega,l)S(\omega,n-l)$ である。任意のスペクトル X の MFCC ベクトルを C_X と表せば次式が成り立つ。

$$C_{H(\omega,l)S(\omega,n-l)} = C_{H(\omega,l)} + C_{S(\omega,n-l)} \quad (3)$$

ここで、クリーン音声 HMM は音声生成の確率モデルであることに着目すれば、$C_{S(\omega,n-l)}$ の分布は HMM の $n-l$ フレームで活動する隠れ状態の出力確率分布で予測できる。$C_{H(\omega,l)}$ は残響特性の MFCC である。従ってこの残響成分は、$n-l$ フレームの隠れ状態の出力確率分布に平均を $C_{H(\omega,l)}$ 加算したものである。分散は変化しない。

続いて、式 (2) よりフレーム n にかかる全残響はそれ以前のフレームに起因する残響成分をスペクトル領域で加算したものである。残響成分のパワーの加法性を仮定すれば、これは過去の各フレームからの残響成分の分布を加法性雑音のためのモデル合成法によって合成することで得られる。本研究では、合成には PMC 法 [1] を用いる。

なお、残響成分の分布を加法することで分散が非常に大きくなるため、ここでは分散は変換対象としない。

3.2 先行音素の情報が未知の場合のモデル変換

本稿で目的とする事前のモデル変換では、音声認識中の先行音素の情報は未知である。そこで、先行音素を確率的に処理する。すなわち、変換する音素に対してそれに先行する音素列のパターンを考え、前節の方法により残響のかかった音素のモデルを得る。これに各々のパターンの出現確率を掛けて重ね合わせることでミクスチャモデルを構成し、その音素の残響モデルとする。ただし、先行音素列のパターンを削減するために以下の仮定をおく。

まず、残響の主要な成分が現在とその直前の音素の 2 音素がカバーする時間内 (およそ 150ms-300ms) のみに由来し、それ以前は無視できると仮定する。音素出現率は音素 bi-gram 確率として単語データから求められる。また、 状態の継続フレーム数は状態遷移確率の逆数である場合のみ考慮する。

以上の仮定によってパターンを削減し、モデル適応を行う。なお、構成されたミクスチャモデルをそのまま認識に用いることも可能ではあるが、多すぎる場合には GMM 推定法などにより少数のミクスチャに統合する。

4 評価

上記提案手法の有効性を検証するため、残響を畳み込んだ音声による特定話者の孤立単語音声認識実験を行った。

4.1 実験条件

評価データには ATR 音声データベース A セットの 655 単語を用い、インパルス応答には RWCP 実環境音声・音響データベースから E2A、E1B、JR1(残響時間はそれぞれ 300ms、310ms、600ms; マイクロフォンと音源の距離は約 2m) を用いて計算機上で畳み込んで残響音声とした。

音声を 16kHz、量子化ビット数 16bit でディジタル化し、フレーム長 25ms、フレーム周期 10ms で分析した。高域強調の定数 $\alpha = 0.97$ とし、窓関数にはハミング窓を用いた。音響特徴量は 0 次を含む 13 次元の MFCC と ΔMFCC の計 26 次元とし、24 個のフィルタバンクによって計算した。音響モデルは 41 音素で 3 状態単混合のモノフォン HMM とし、評価データと同一話者の 2620 単語で学習した。学習したクリーン音響モデルに、評価音声に畳み込んだものと同一のインパルス応答を用いて提案手法のモデル変換を施した。ただし、変換時に用いる各音素列の出現確率は認識対象の単語から事前に求めて与えた。デコーダには Julian を用いた。

比較手法として、フレーム外の残響に対応できているかを検証するため、インパルス応答を最初の 1 フレームだけ与えて、モデル変換を行った。他のフレームの残響は存在しないと考える。これは観測フレーム外の残響を考慮せず、直接音の伝送路歪みのみに対応したモデル変換であると考えられる。

また、インパルス応答の変化に対するロバスト性を検証するため、評価音声に畳み込んだものとは若干異なるインパルス応答を与えてモデル変換も行った。これには、評価音声と同一環境だが約 30cm 離れたマイクロフォンで観測したインパルス応答を用いた。

表 1: 提案手法による残響音声の単語認識率（%）

残響特性	baseline	short	proposed	mismatched	retrained
E2A	74.81	81.68	86.87	85.34	92.98
E1B	63.36	73.28	75.42	72.37	85.34
JR1	38.02	54.35	60.76	55.88	77.71

4.2 実験結果

結果を表 1 に示す。認識に用いたモデルはそれぞれ、クリーン音声モデル (baseline)、1 フレームのみの残響で変換したモデル (short)、提案手法を施したモデル (proposed)、評価音声と異なるインパルス応答を与えて提案手法を施したモデル (mismatched)、評価音声と同一の残響特性下の音声で学習したモデル (retrained) である。

4.3 考察と今後の展望

表 1 から提案手法による認識率の向上が確認されたが、retrained に比べ低く、改善の余地がある。残響成分を含めたことによる short からの改善効果は環境によって異なり、また単純に残響時間との関連はみられない。このことから、認識率が低いのは考慮に入れた残響が短かったというよりも、残響推定の精度や 3.2 節で行ったパターン削減による偏りが原因であると考えられる。これを解決するために、例えばコンテキスト依存 HMM に対して本手法を適用することが考えられる。コンテキスト依存 HMM では、ある音素に対してその直前にある音素が決定できるため、直前音素の内容を統計的に扱う必要はない。mismatched における認識率の低下は低く、インパルス応答の変化に対するロバスト性が確認された。

5 おわりに

本稿では、残響に頑健な音声認識手法として、 クリーン音声のモデルと残響特性を与えて、残響環境に適応させた音響モデルを作成する手法を提案した。変換する音素に対してその直前にある音素列の可能性を場合分けし、それぞれの場合で先行フレームからの残響成分をモデル合成して残響モデルを求めた。そして残響モデルを音素列の出現確率によって重ね合わせて、変換結果とした。残響下音声の特定話者孤立単語音声認識実験により認識率の向上が確認できたが、まだ改善の余地がある。

今後の課題としては、残響分布推定の精度向上や不特定話者コンテキスト依存 HMM などを検討したい。

参考文献

[1] M. J. F. Gales, S. J. Young: “Robust Continuous Speech Recognition Using Parallel Model Combination,” IEEE Trans. Speech and Audio Process. 4, pp. 352-359, 1996.

[2] H. Hermansky, N. Morgan: “RASTA Processing of Speech,” IEEE Trans. on Speech and Audio Process. 2, pp. 578-589, 1994.

[3] 滝口哲也, 中村哲, 鹿野清宏: “加法性雑音、伝達特性による歪みを受けた音声の HMM 合成による認識,” 日本音響学会 1995 年秋季研究発表会講演論文集 1-2-2, pp. 3-4, 1995.

[4] 広林茂樹, 野村博昭, 小池恒彦, 東山三樹夫: “パワーエンベロープ伝達関数の逆フィルタ処理による残響音声の音源波形回復,” 電子情報通信学会論文誌, Vol. J81-A, no. 10, pp. 1323-1330, 1998.