E-045

RSF とマルチコンディション HMM による 雑音ロバスト孤立単語音声認識
Robust Isolated Word Speech Recognition by RSF and Multi-Condition HMM

千歳 広大† 早坂 昇‡ 吉澤 真吾† 宮永 喜一†
Koudai Chitose Noboru Hayasaka Singo Yoshizawa Yoshikazu Miyanaga

1. まえがき

音声認識技術も実用化の段階に入り，音声入力インターフェイスを搭載した機器も開発されている．しかし実環境では環境雑音や残響音の影響を大きく受け，認識率が低下してしまう．そのため雑音対策手法に関する研究が長く行われている [1]-[6].

音声信号処理における雑音低減手法としては，パワースペクトルの変調スペクトル及び対数パワースペクトルの変調スペクトル[2]の両方に，時間軌跡に対してフィルタリング処理を行い，音声の強調，乗法性雑音の除去及び加法性雑音の抑制を行うランニングスペクトルフィルタリング法がある[3].

また音声認識に用いる音響モデルの雑音対策手法としては，複数の雑音環境下に頑健なマルチコンディション HMM（Multi-HMM）[4],[5],[6]が提案されている．

本稿では先ず RSF 及びマルチコンディション HMM の簡単な説明を行う．次に両者を融合した新しいロバスト音声認識方式について述べ，それを用いた孤立単語音声認識実験を行い，提案法の有用性を示す．

2. 提案法

2.1 ランニングスペクトルフィルタリング（RSF）

2.1.1 変調スペクトル

音声信号は性質が常に変化している非定常信号であるが，短時間区間では定常的な信号とみなすことが出来る．フレームによる短時間スペクトル分析を行い，フレーム周期毎に求められるスペクトルの時系列データをランニングスペクトルと呼ぶ．このランニングスペクトルをフーリエ変換したものを変調スペクトルと呼ぶ．

ランニングスペクトルには対数変換する前のパワースペクトルと，対数変換した後のパワースペクトルがある．どちらの変調スペクトルも 7Hz 以下の部分のみを用いても，認識性能は低下しないという特徴を持っている．

加法性雑音はパワースペクトルに対する変調スペクトルの低域に多く分布し，また乗法性雑音は対数パワースペクトル領域において加算の形で存在し，時間変化がほとんどないため，変調スペクトルの 0Hz 付近に分布する．

2.1.2 ランニングスペクトルフィルタ

変調スペクトルの特徴を考慮し，ランニングスペクトルフィルタとして，パワースペクトル及び対数パワースペクトルの時間軌跡に対しフィルタリングを行う．

パワースペクトルに対する変調スペクトルの低域には多くの加法性雑音が存在しているが，他の帯域にも広く分布する．また加法性雑音が非定常な場合，低域以外の部分にも強く影響する．そこで 7Hz 以下を強調するようなローパスフィルタリングを施すことにより，音声の特徴成分を強調し，7Hz 以上の加法性雑音の影響を抑えることが可能になる．

対数パワースペクトルの変調スペクトルに対するフィルタリングでは，低域に乗法性雑音が集中しているため，バンドパスフィルタを用いることにより乗法性雑音の除去を行っている．

ランニングスペクトルフィルタは FIR 型で設計されており，計算量は多いが位相歪みが発生しないため，安定して雑音の抑制が可能になる．

2.2 Multi-HMM

音声認識を行う場合，一般的にその音響モデルには雑音のない音声データを学習時に用いる．しかし，実環境下では雑音や残響音による影響が存在するため，認識率が低下してしまう．よって，学習時に様々な雑音を重畳した音声データで学習を行い，雑音音声のモデル化を行う．これを Multi-HMM と呼ぶ．このモデルを用いる事により，実環境下における様々な雑音に対応できる HMM を実現することが目的とされている．一般的には，この HMM を実現するためには，多くの雑音源を想定し，様々な SNR を仮定する．そのため，Multi-HMM が未知の雑音に対しても頑健となるには，多くのデータを必要とし，学習に要する時間が多くなってしまう．さらに，想定される SNR によっては認識性能を劣化させてしまい，従来では，認識時において，雑音の種類をある程度特定化し，その後複数作られた Multi-HMM の中から，最適な Multi-HMM を選んで認識するような，限定環境下における認識や，スペクトルサブトラクション(SS)法等の雑音ロバスト処理を活用することにより，音声信号に対する雑音の影響をできるだけ少なくした上で Multi-HMM を利用して認識性能を上げるなどの，様々な工夫がされてきた．

本稿では学習に要する時間の削減及び認識率の向上を目指し，雑音ロバスト処理として，高い性能を実現できる RSF を，学習と認識に適用することを前提に，学習音声データとして，NOISEX-92 から複数の雑音を用いて SNR 比 0dB – 20dB の雑音音声を生成し，さらにクリーン音声データの合計 6 種類の学習用音声データを合成した．

3. 孤立単語音声認識実験

3.1 実験条件

提案法を評価するために孤立単語音声認識実験を行った．量子化ビット幅 16bit，サンプリング周波数 11.025kHz の 142 個の音声データを用いた．学習には男性話者 40 名

†北海道大学情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
‡株式会社レイトロン RayTron,INC

1 人 3 回発声分を用い，学習に用いなかった不特定話者 30 名 1 人 1 回発声分を認識データとして用いた．また，NOISEX-92 にある 15 種類の雑音を SNR 比 0dB, 10dB, 20dB で人為的に付加し，音声認識の実験データとして用いた．

今回の実験に関する分析条件を表 1 に，HMM 初期値を表 2 に示す．また，HMM モデルは単語モデルの left-to-right モデルを使用した．

実験は，雑音のない環境の音声のみで出来た HMM と，提案法であるマルチコンディション HMM の 2 種類で比較した．

表 1　分析条件

Window length	23.2ms(256point)
Shift length	11.6ms(128point)
Preemphasis	$1\text{-}0.9688z^{-1}$
Feature vectors	MFCC12, ΔMFCC12, ΔΔMFCC12

表 2　HMM 初期値

状態数	32 個
混合数	1 個
初期状態分布	$\pi_0=1$, $\pi_i=0$
状態遷移確率	$a_{ii}=a_{ii+1}=0.5$, $a_{ij}=0$
学習繰り返し回数	繰り返し前の尤度と繰り返し後の尤度の平均の差が 0.5 未満になると終了．

3.2 実験結果

表 3 及び表 4 に実験結果を示す．高 SNR 時における Car interior noise の認識率は，クリーン音声だけを利用して学習した HMM による認識結果が，97.00 に対して，マルチコンディション HMM を利用した時の認識率が，96.50 となり，若干認識率が低下している．他の条件や雑音に関しては，認識率の上昇がみられる．特に低 SNR 環境下における認識性能が大きく向上しており，実環境下における音声認識において，提案法が有用であると考えられる．

4. まとめ

本報告では，雑音除去手法の 1 つである RSF と，雑音重畳データを学習に用いるマルチコンディション HMM を併用する手法が，雑音除去に対して有用であるか検証した．

提案法では，ほぼ全ての雑音に対して認識率が上昇し，特に低 SNR 環境下における認識性能が大きく向上した．

今後，更に雑音環境下における認識率の向上を目指し、新たな手法を検討していく予定である．

表 3　雑音のない環境のみの HMM を用いた場合の認識率

Noise name\SNR	0dB	10dB	20dB
White noise	22.07	61.73	85.47
Pink noise	22.43	70.03	90.67
HF channel noise	21.37	67.17	88.10
Speech babble	9.33	59.33	88.70
Factory floor noise 1	11.50	63.87	90.60
Factory floor noise 2	50.13	87.40	95.47
Jet cockpit noise 1	15.80	64.73	89.60
Jet cockpit noise 2	19.47	63.70	87.10
Destroyer engine room noise	24.43	74.47	91.97
Destroyer operation room noise	15.47	67.67	91.60
F-16 cockpit noise	26.73	75.90	92.87
Military vehicle noise	79.67	92.23	95.57
Tank noise	51.83	88.10	95.40
Machine gun noise	54.20	70.20	82.63
Car interior noise	94.67	96.47	97.00
Average	39.14	73.53	91.01

表 4　マルチコンディション HMM を用いた場合の認識率

Noise name\SNR	0dB	10dB	20dB
White noise	56.76	85.90	93.17
Pink noise	56.76	88.03	94.33
HF channel noise	49.83	85.80	93.70
Speech babble	30.83	87.73	95.47
Factory floor noise 1	35.97	87.73	94.70
Factory floor noise 2	76.43	93.13	96.07
Jet cockpit noise 1	45.97	84.90	93.97
Jet cockpit noise 2	52.67	85.13	92.87
Destroyer engine room noise	56.90	87.50	94.47
Destroyer operation room noise	44.20	88.77	95.13
F-16 cockpit noise	59.67	88.90	94.87
Military vehicle noise	88.03	93.63	95.57
Tank noise	76.67	92.70	95.53
Machine gun noise	70.97	84.43	92.20
Car interior noise	94.93	96.27	96.50
Average	64.07	88.70	94.46

参考文献

[1]Lawrence Rabiner,Biing-Hwang Juang,古井 貞熙, “音声認識の基礎(下)”, NTT アドバンステクノロジ株式会社, pp.102-182 (1995).

[2]金寺 登, 荒井 隆行, 船田 哲男, “変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識”, 信学論, Vol.J84-D-II, No.7, pp.1261-1269 (2001).

[3]早坂 昇, 和田 直哉, 宮永 喜一, 畑岡 信夫, “ランニングスペクトルフィルタを用いた雑音にロバストな音声認識”, 信学会, 信学技報, CAS2003-6, pp.31-36 (2003).

[4]Davit Pearce and Hans-Gunter Hirsch, “The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition System Under Noisy Conditions”, Proc.of ICSLP2000, Vol.4, pp.29-32 (2000).

[5]福士 ななこ, 加藤 正治, 小坂 哲夫, 好田 正紀, “ETSI 標準フロントエンドを用いた雑音下音声認識の検討”, 信学会, 信学技報, SP2004-11, pp.7-12 (2004).

[6]西亀 健太, 渡部 晋治, 西本 卓也, 小野 順貴, 嵯峨山 茂樹, “複数残響特性下の音声を単一モデル学習に用いた未知残響環境に頑健な音声認識の検討”, 信学会, 信学技報, SP2008-8, pp.43-48 (2008).