社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（28）
2005／12／22

Model-Based Wiener Filterによる雑音の種類に頑健な音声認識

荒川　隆行　　辻川　剛範　　磯谷　亮輔

NECメディア情報研究所

E-mail: t-arakawa@cp.jp.nec.com, tujikawa@cb.jp.nec.com, r-isotani@bp.jp.nec.com

あらまし　雑音の種類に頑健な音声認識手法である Model-Based Wiener Filter 法を提案する．本手法は，信号処理的手法である Wiener Filter と，音声 GMM による MMSE 推定法を組み合わせたものである．AURORA2-J タスクを用いて評価を行ったところ，ETSI の提唱する Advanced Front-End と較べて単語正解精度が平均値で同程度，雑音の種類によるばらつきでは 3 分の 1 程度となり，提案法が雑音の種類に頑健に動作することが確かめられた．

Model-Based Wiener Filter for Noise Robust Speech Recognition

Takayuki ARAKAWA, Masanori TSUJIKAWA, and Ryosuke ISOTANI

NEC Media and Information Research Laboratories

E-mail: t-arakawa@cp.jp.nec.com, tujikawa@cb.jp.nec.com, r-isotani@bp.jp.nec.com

Abstract　We propose a new approach for noise robust speech recognition, Model-Based Wiener Filter. This method takes three steps to estimate clean speech signals from noisy speech signals. The first step is the spectral subtraction (SS). Since the SS averagely subtracts noise components, the estimated speech signals often include distortion. In the second step, the distortion caused by SS is reduced using the minimum mean square error estimation for a Gaussian mixture model. In the final step, the Wiener Filtering is performed with the decision-directed method. Experiments are conducted using the AURORA2-J database. The results show that the proposed method performs as well as the ETSI advanced front-end in average and the variation range of the recognition accuracy according to the kind of noise is about one third, which demonstrates the robustness of the proposed method.

1. はじめに

モバイル端末や車載端末用の入力手段として，背景雑音に頑健な音声認識技術が期待されている．このような要望に対し，Wiener Filter を用いた方法である ETSI の提唱する Advanced Front-End (AFE) [1] や，音声 Gaussian Mixture Model (GMM) を用いた MMSE 推定法 [2], [3] が注目されている．前者の方法は，SNR の平滑化を強く行うことで，特に定常雑音で高い性能が得られる．これに対し後者の方法は，音声のモデルがどのように雑音の影響を受けるかを考慮し雑音の低減を行うため，雑音の種類に依らず頑健に動作するが，音声モデルを雑音に適応させる処理に多くの計算量を必要とする．そこで両者の特徴を組み合わせ，少ない計算量で，様々な種類の雑音に対し高い性能となる Model-Based Wiener Filter (MBW) 法を提案する．

2. Model-Based Wiener Filter 法

以下に，MBW 法のアルゴリズムについて述べる．

Ⅰ．入力された雑音混じりの音声スペクトル $\boldsymbol{X}(t)$ から雑音平均スペクトル $\overline{\boldsymbol{N}(t)}$ を推定する．$\overline{\boldsymbol{N}(t)}$ の推定には入力信号の始めの無音区間の平均値などを用いる．$\boldsymbol{X}(t), \overline{\boldsymbol{N}(t)}$ は，周波数スペクトルを成分に持つベクトル，t はフレーム番号である．

Ⅱ．スペクトル減算法 (SS 法) を行い，仮推定音声 $\overline{\boldsymbol{S}_{\rm SS}(t)}$ を求める．

$$\overline{\boldsymbol{S}_{\rm SS}(t)} = \max(\boldsymbol{X}(t) - \overline{\boldsymbol{N}(t)}, \alpha \boldsymbol{X}(t)) \tag{1}$$

ここで α はフロアリングパラメータである．

Ⅲ．仮推定音声をケプストラムに変換する．

$$\overline{\boldsymbol{C}_{\rm SS}(t)} = {\rm DCT}\left[\log(\overline{\boldsymbol{S}_{\rm SS}(t)})\right] \tag{2}$$

DCT $[\cdot]$ は離散コサイン変換を示す．$\overline{\boldsymbol{C}_{\rm SS}(t)}$ はケプストラム次元を成分に持つベクトルである．

Ⅳ．音声期待値を求める．

あらかじめクリーンな音声で学習した GMM を用いる．K 混合の GMM は式 (3) のように表す事ができる．

$$P(\boldsymbol{C}) = \sum_{k=1}^{K} P(k) P(\boldsymbol{C}|k) \tag{3}$$

ここで，$P(k)$ は k 番目のガウス分布の混合重み（事前確率），$P(\boldsymbol{C}|k) = \mathcal{N}(\boldsymbol{C}|\mu_k^{\rm cep}, \Sigma_k^{\rm cep})$ は音声 $\boldsymbol{C}$ に対する k 番目のガウス分布の尤度である．このとき，仮推定音声 $\overline{\boldsymbol{C}_{\rm SS}(t)}$ に対する k 番目のガウス分布の事後確率は式 (4) となる．

$$P(k|\overline{\boldsymbol{C}_{\rm SS}(t)}) = \frac{P(k) P(\overline{\boldsymbol{C}_{\rm SS}(t)}|k)}{\sum_{k=1}^{K} P(k) P(\overline{\boldsymbol{C}_{\rm SS}(t)}|k)} \tag{4}$$

この事後確率を用いて音声期待値を求める．

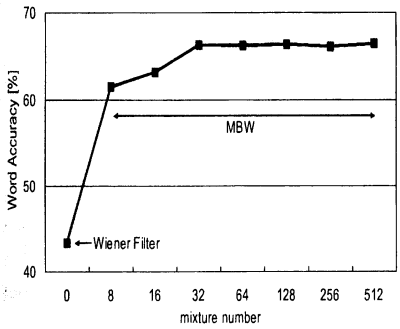

図 1　混合数と単語正解精度
(レストラン雑音 SNR=5dB)

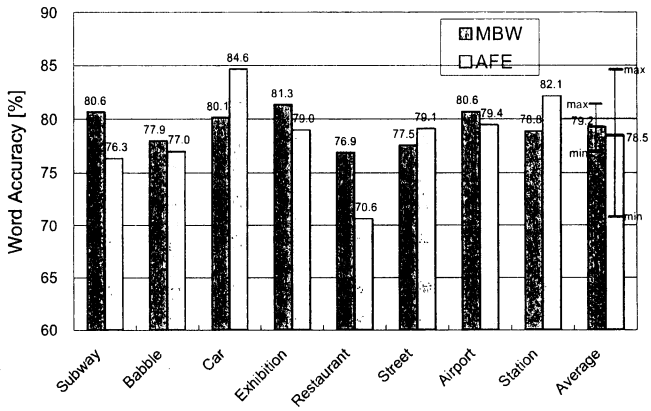

図 2　雑音の種類と単語正解精度

$$\overline{\langle \boldsymbol{S}(t) \rangle} = \exp\left[\sum_{k=1}^{K} P(k|\overline{\boldsymbol{C}_{\mathrm{SS}}(t)})\boldsymbol{\mu}_{\boldsymbol{k}}^{\log}\right] \tag{5}$$

ここで, $\boldsymbol{\mu}_{\boldsymbol{k}}^{\log}$ はケプストラム量ではなく, 対数スペクトルの量として予め計算しておいたものである.

$$\boldsymbol{\mu}_{\boldsymbol{k}}^{\log} \equiv \mathrm{IDCT}[\boldsymbol{\mu}_{\boldsymbol{k}}^{\mathrm{cep}}], \tag{6}$$

ここで IDCT $[\cdot]$ は逆コサイン変換を示す.

V.　ウィナーゲインを求める.

$$\boldsymbol{W}(t) = \frac{\boldsymbol{\eta}(t)}{\boldsymbol{\eta}(t)+1} \tag{7}$$

上式の $\boldsymbol{\eta}(t)$(平滑化事前 SNR) は, 次式のように求める [4].

$$\boldsymbol{\eta}(t) = \beta\boldsymbol{\eta}(t-1) + (1-\beta)\frac{\overline{\langle \boldsymbol{S}(t) \rangle}}{\overline{\boldsymbol{N}(t)}} \tag{8}$$

ここで β は平滑化パラメータである.

VI.　最終的なクリーン音声推定値を求める.

$$\overline{\boldsymbol{S}(t)} = \boldsymbol{W}(t)\boldsymbol{X}(t) \tag{9}$$

式 (5) の結果をそのまま推定値とすると, 特に GMM の混合数が少ないとき, 元の入力音声の持つ情報が過剰に失われる. このためウィナーゲインを算出し, 式 (9) のように音声の推定値を求める. また Ⅲ.～VI. の手順をくり返し行うことでさらに精度の高い推定を行うことが出来る.

3. 評価実験

3.1 評価データ

評価データには AURORA-2J タスク（日本語連続数字読み上げ) を使用した [5]. 用いたのはテストセット A およびテストセット B である. 学習はクリーンコンディションで行った.

3.2 評価条件

音響分析の条件は, 標本化周波数を 8kHz(16bit), 特徴量を 13 次 MFCC(0 次含む) + Δ + ΔΔ の計 39 次元とした. 雑音抑圧用に使用する GMM には 13 次元 MFCC(0 次含む) を用いた. また GMM の学習には音響モデル作成時と同じ学習データを使用している. 式 (1) で用いるフロアリングパラメータ α は 0.1, 式 (8) で用いる平滑化パラメータ β は 0.98 とした. また, 前章の手順 Ⅲ.～VI. を二回くり返して用いている.

3.3 評価結果

レストラン雑音 (SNR=5dB) について, 提案法における GMM の混合数を変えたときの単語正解精度を図 1 に示す. 図の左端 0 混合は Wiener Filter 単体での性能である. 混合数を増やすと性能が向上しており, 提案法の効果が見られる. この結果を踏まえて, 以下の実験では混合数を 256 とする.

雑音の種類を変えた結果を図 2 に示す. 縦軸は単語正解精度 (SNR = 20dB～0dB の平均値), 横軸は雑音の種類である. また, 右端に全体の平均値および最大値と最小値を示した. 比較のために, 提案法 (MBW) に加えて ETSI Advanced Front-End (AFE) [1] の結果を示す. 提案法は AFE に較べて平均性能で同程度, 雑音の種類によるばらつきでは 3 分の 1 程度となり, 提案法が雑音の種類に頑健に動作すると言える.

4. おわりに

信号処理的手法である Wiener Filter と, 音声 GMM による MMSE 推定法を組み合わせて用いた手法である Medel-Based Wiener Filter を提案し, 評価を行った. 提案法は従来法に較べて雑音の種類に依らず頑健に動作することを確認した. 今後は加法性雑音だけではなく, 乗算性雑音や突発雑音に対しても頑健な手法となるよう改良を行う予定である.

文　献

[1] ETSI ES 202 050 v1.1.1, 2002.
[2] J.C.Segura *et al.*, EuroSpeech'01, Vol.I, pp.221-224(2001).
[3] 藤本雅清他, 情報処理学会研究報告, SLP-47-16(2003).
[4] Y. Ephraim *et al.*, IEEE Trans. ASSP 32, pp.1109-1121, 1984.
[5] S. Nakamura, *et al.*, ASRU2003, pp.619-623, 2003.