研究開発

Model-Based Wiener Filterによる雑音下音声認識

荒川隆行・辻川剛範

要　旨

音声認識の性能は、雑音が存在する環境では著しく劣化することが知られています。このような雑音による影響を軽減するために、雑音が音声に較べてゆっくり変動するという雑音の知識と、音声らしさをモデル化した音声の知識を用いることで、雑音のみを取り除き音声を強調するModel-Based Wiener Filter法について紹介します。

キーワード

●音声認識　●耐雑音　●Wiener Filter

1. はじめに

音声認識の性能は雑音の影響によって著しく劣化してしまうため、実運用に当たっては耐雑音技術が重要となります。雑音による性能劣化の原因は、音響モデルを作成する際に使用する音声データと、実運用で使用する音声データとの間で雑音環境のミスマッチが生じるためです。人の発声する音声は、のどを通ってくる空気の振動に対し、口の形状や舌の位置などを使って周波数に変化を与えることで作られます。このような音の変化を音素ごとに確率密度関数で表したものが音響モデルです。音素とは音の区分の最小単位です。母音や子音、促音、撥音などで区分する単位が用いられます。実環境では、人が発声した音声に加えて、周囲の雑音の周波数特性や音声が伝達してくるまでの伝送経路の周波数特性などの影響でミスマッチが生じます。このようなミスマッチを解消するために多くの手法が提案されています[1~5]。

音響モデルと実運用で使用する音声データとのミスマッチを改善する方法は大きく分けて2通りの手法が知られています。1つは実運用で使用する音声データ側を補正する方法です[1,2]。実運用で使用する音声データ中の雑音成分を推定し、除去することによりクリーンな音声に近づけます。もう1つは音響モデル側を雑音環境に適応させる方法です[3]。前者は比較的少ない計算量で雑音を除去できる利点と、雑音成分の時間変化に追従する方法[1]と組み合わせることができる利点があります。後者は音声を構成する音素ごとに確率密度関数が雑音によってどのように変化するかを扱うことができます。

本稿で紹介するModel-Based Wiener Filter（MBW）法[4,5]は、前者の雑音を除去する方法に対して、音声モデルの知識を用いることで雑音の種類によらず不要な雑音成分を取り除き、頑健に音声を強調する手法です。

2. Model-Based Wiener Filter法

以下、MBW法の原理について説明します。

時刻kにおいてマイクロホンから入力された信号 x（k）は、音声 s（k）と、加算性雑音 n（k）と、乗算性雑音 h とで表されます。

$$x(\mathrm{k}) = h \cdot s(\mathrm{k}) + n(\mathrm{k}) \qquad (1)$$

乗算性雑音は、マイクロホン特性や話者性を表します。発声の前後及び発声中にこの乗算性雑音はほとんど変化しないため一定として考えます。短時間フーリエ変換を行うと、周波数帯域ごとに式（2）が得られます。

$$X(\mathrm{t},\mathrm{f}) \approx H \cdot S(\mathrm{t},\mathrm{f}) + N(\mathrm{t},\mathrm{f}) \qquad (2)$$

ここでtはフレーム番号、fは周波数帯域を示します。X（t,f）は入力信号のスペクトルパワー、Hは乗算性雑音のスペクトルパワー、S（t,f）は音声のスペクトルパワー、N（t,f）は加算性雑音のスペクトルパワーをそれぞれ示します。式（2）では、音声と加算性雑音の間の位相の差が平均的に0であり、無視できるものと近似しています。**図1**は音声のスペクトルパワーと加算性雑音のスペクトルパワーの足し合わせで、入力信号のスペクトルパワーが得られることを示したものです。各グラフの横軸は周波数[Hz]、縦軸はスペクトルパ

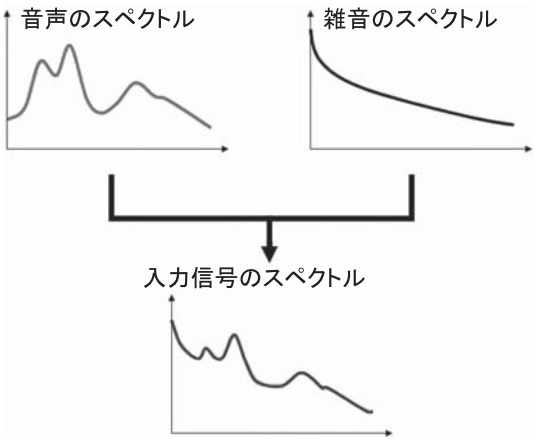

図1　音声のスペクトルと加算雑音のスペクトル

ワーの対数値[dB]を示します。

図1で見て取れるように加算性雑音の影響でスペクトルパワーの形状が大きく変化しています。特にスペクトルの谷が雑音に埋もれており、この谷構造を復元することが重要になります。入力信号の時系列から、谷構造を復元し音声成分のみを得るためには、加算性雑音の成分と乗算性雑音の成分を推定、除去する必要があります。乗算性雑音は一定であるため、ケプストラム平均値正規化法[6]を用いることにより除去することが可能です。しかしながら、音声パワー、加算性雑音パワーは共に時間変化するために、式(2)からだけでは雑音成分を除去することができません。MBW法では、加算性雑音の時間変化は音声に較べてゆっくりであるとの雑音の知識と、音声のスペクトル形状はあらかじめおおよその形が分かっているとの音声の知識を用いることで、雑音成分を除去し、音声成分のみを取得します。

アルゴリズム

MBW法では音声の知識として、あらかじめクリーンな音声で作成した混合ガウス分布（GMM）を用意します。GMMは音声をケプストラム（スペクトルの包落形状を表現する特徴量）の形に変換し、その出現頻度を表す確率密度関数を複数のガウス分布の重ね合わせで表したものです。

$$p(\mathbf{V}) = \sum_{k} p(\mathrm{k}) f(\mathbf{V};\mathrm{k}) \qquad (3)$$

ここで、V（t）はtフレームの音声から計算されたケプストラムのベクトルです。f（V;k）は、k番目のガウス分布を表しp（k）はk番目のガウス分布が選択される確率（事前確率）を表します。

以下、MBW法のアルゴリズムについて説明します。

1) まず、入力信号のパワーX（t,f）から、雑音平均スペクトルN（t,f）を推定します。雑音の推定にはSNRに応じた重み付き平均を用いる方法[1]を用います。

2) 次に、スペクトル減算法を用いて仮推定音声$\mathrm{S}_{\mathrm{temp}}$（t）を求めます。

$$\mathbf{S}_{\mathrm{temp}}(\mathrm{t}) = \max\left(\mathbf{X}(\mathrm{t}) - \mathbf{N}(\mathrm{t}), \alpha\mathbf{X}(\mathrm{t})\right) \qquad (4)$$

ここで、太字はベクトルを表します。αはフロアリングパラメータです。max（,）はどちらか大きい値を取る演算を示します。

3) 次に、仮推定音声をケプストラムに変換します。

$$\mathbf{V}_{\mathrm{temp}}(\mathrm{t}) = \mathrm{DCT}[\log \mathbf{S}_{\mathrm{temp}}(\mathrm{t})] \qquad (5)$$

ここで、DCT[ ]は離散コサイン変換を表します。

4) 次に、式（3）の確率密度関数と仮推定音声を用いて、式（6）で示される事後確率を算出します。

$$p(k \mid \mathbf{V}_{\mathrm{temp}}(\mathrm{t})) = \frac{p(\mathrm{k}) f(\mathbf{V};\mathrm{k})}{\sum_{k'} p(\mathrm{k}) f(\mathbf{V};\mathrm{k}')} \qquad (6)$$

事後確率は、仮推定音声が得られたときにk番目のガウス分布が選択される確率を示します。

5) 次に、事後確率を用いて補正された推定音声を算出します。補正された推定音声は式（7）で求められます。

$$\mathbf{S}_{\mathrm{MMSE}}(\mathrm{t}) = \sum_{k} \boldsymbol{\mu}_{S}\, p(k \mid \mathbf{V}_{\mathrm{temp}}(\mathrm{t})) \qquad (7)$$

ここで、μSはk番目のガウス分布から出力される音声パワースペクトルの平均値です。

6) 次に、式（8）を用いて補正された推定音声からウィナーゲインを算出します。

$$W(\mathrm{f},\mathrm{t}) = \frac{S_{\mathrm{MMSE}}(\mathrm{f},\mathrm{t})}{S_{\mathrm{MMSE}}(\mathrm{f},\mathrm{t}) + N(\mathrm{f},\mathrm{t})} \qquad (8)$$

7) 次に、ウィナーゲインを元の入力音声に乗算することにより音声の再推定値を得ます。

$$S_{WF}(\mathrm{f},\mathrm{t}) = W(f,t)\cdot X(f,t) \qquad (9)$$

8) 最後にホワイトノイズを付加し、消し残しの雑音を正規化します。

$$S(\mathrm{f},\mathrm{t}) = S_{\mathrm{WF}}(\mathrm{f},\mathrm{t}) + \mathrm{flr}. \qquad (10)$$

図2 にMBW法の動作原理を示します。式（4）で求めた仮推定音声は、雑音推定時の誤差や、位相の差を無視したことによる誤差のために、元の音声からずれることがあります。特にスペクトルパワーの形状の谷の部分では誤差の影響が現れやすい傾向があります。式（6）で求めた事後確率とは、入力音声とGMMの各分布とのマッチングの割合のことであり、式（7）ではこの事後確率を用いて、より標準的な音声に近づくように補正しています。このようにすることで失われてしまったスペクトルパワーの谷構造を復元することができます。更にウィナーフィルターの形にすることにより、位相の差を無視したことによる誤差を軽減することができます。最後にホワイトノイズを付加することで、消し残しの雑音を目立たなくするとともに、非音声区間のスペクトル形状を音響モデル中の無音モデルに近づけることができ、湧き出し誤りを削減することができます。

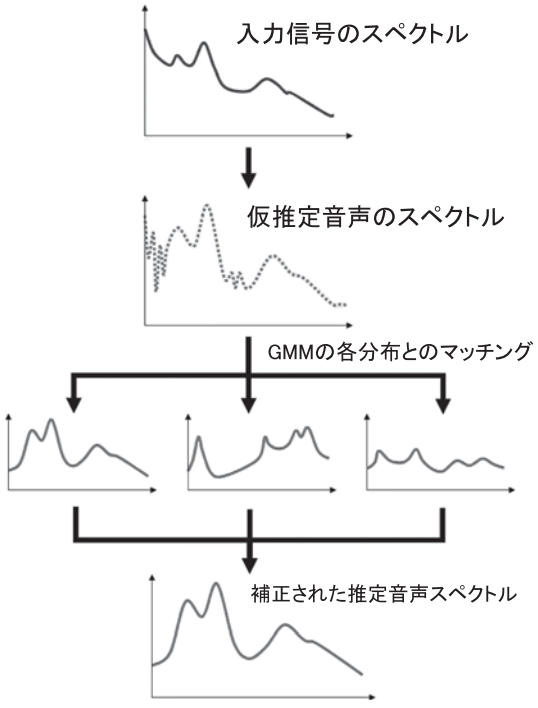

図2　MBW法の動作原理

3. 音声認識評価

次に、MBW法を用いた音声認識評価について説明します。

(1) 学習データ、評価データ

学習データ及び評価データにはIPSJ-SIG SLP雑音下音声評価ワーキンググループが配布しているAURORA-J タスク（日本語連続数字読み上げ）を使用しました[7]。音響モデルの作成にはクリーンな音声のみを使用しました。

(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）とその1次差分量及び2次差分量の計39次元を用いました。MBW法に用いるGMM には13 次元のMFCC（0次含む）を用いました。またGMM の作成には音響モデル作成時と同じデータを使用しました。式（4）で用いるフロアリングパラメータ α は0.1を使用しました。

(3) 評価結果

MBW法におけるGMMの混合数を変えたときの認識率（単語正解精度）を **図3** に示します。ここではレストラン雑音を用いました。レストラン雑音は非定常な雑音が含まれるために、比較的MBW法の効果が現れやすい雑音であると考

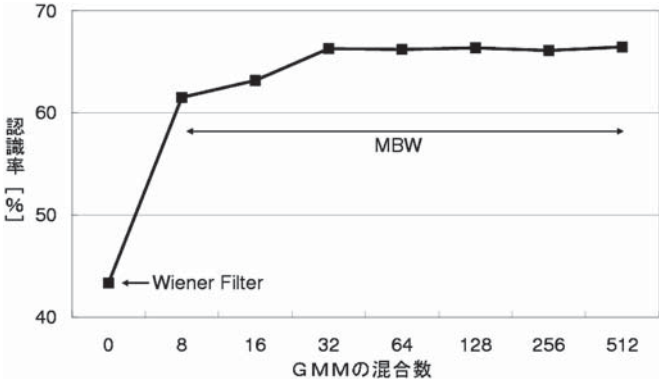

図3　GMMの混合数と認識率

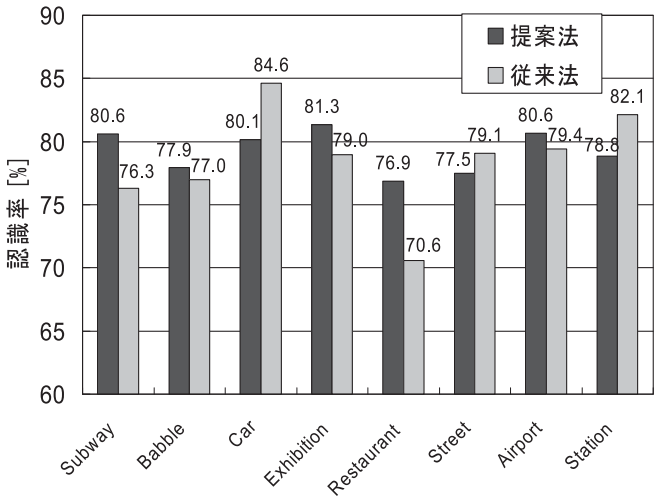

図4　雑音の種類と認識率

えられます。図3の左端0混合はWiener Filter単体での性能です。混合数を増やすにつれて性能が向上しており、MBW法の効果が見られます。この結果を踏まえ、以下混合数は256を用います。

雑音の種類を変えた結果を**図4**に示します。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す。比較のために、提案法（MBW）に加えて従来法であるETSI Advanced Frontend（AFE）[1]の結果を示します。AFEはWiener Filterを2回に分けて行うことで、雑音の消し残しを小さくすることができ、特に定常雑音において性能が高い方法です。提案法は従来法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いえます。

4. まとめ

雑音の知識に加えて、音声の知識を用いることで、雑音の種類によらず頑健に音声の強調を行えるModel-Based Wiener法について紹介しました。ここでは音響モデルをクリーンな音声を用いて学習する方法のみを説明しましたが、本手法は、複数の雑音環境で収録された音声を用いて学習するマルチコンディション学習を行うことにより更に高い性能が得られることが確かめられています[4]。今後この手法を使って音声認識の適応範囲が広がることが期待されます。

参考文献

1) 加藤正徳ほか，“重み付き雑音推定とMMSE STSA法に基づく高音質雑音抑圧.” 2004年
2) ETSI ES 202 050 v1.1.1，“Distributed speech recognition; front-end feature extraction algorithm; compression algorithm.” 2002年
3) P. J. Moreno et al.，“A vector Taylor series approach for environ - ment-independent speech recognition.” 1996年
4) 荒川隆行ほか，“Model-Based Wiener Filter による雑音下音声認識.” 2005年
5) 辻川剛範ほか，“Model-Based Wiener FilterとMulti-Condition学習の併用による車内音声認識.” 2008年
6) 黒岩眞吾ほか，“最尤状態系列を用いた実時間ケプストラム平均値正規化の検討.” 1998年
7) 山本一公ほか，“AURORA-2J/AURORA-3Jデータベースとのその評価ベースライン.” 2003年

執筆者プロフィール

荒川 隆行
共通基盤ソフトウェア研究所
主任

辻川 剛範
共通基盤ソフトウェア研究所
主任