マイクロホンアレー入力の周波数領域での幾何学的処理による雑音環境下の音声認識*

◎岡嶋崇（東大・工）　鎌本優　西本卓也　嵯峨山茂樹（東大・情報理工）

1　はじめに

実環境での自動音声認識では、周囲雑音や残響が認識性能の劣化の原因となる。マイクロホンアレーを用いることにより、対象音と雑音の空間的情報を利用し、それらの影響を抑制し、遠隔発話音声の認識性能を向上させることができる。

マイクロホンアレーについては様々な技術が研究されてきている。基本的な方法である Delay-and-Sum (DS) は学習を必要としないが、その性能は十分とはいえない。Griffith-Jim や AMNOR などの適応フィルタ型マイクロホンアレーでは、予め無音声区間を入力し学習させることが必要である [1]。しかし、雑音や残響が時々刻々変化する環境では、学習による環境への追従が間に合わず、性能が低下することがある。

また、それら従来の手法は信号波形の推定を主な目的としているが、音声認識に通常用いられる MFCC のような音響特徴量を計算するためには信号の短時間スペクトルを推定すれば十分である。

そこで本稿では、音声認識の性能向上を目指し、マイクロホンアレーを用い、周波数領域での幾何学的な考察に基づいた高性能かつ学習の必要のない短時間スペクトル推定の方法について検討したので報告する。

2　周波数領域での幾何学的処理

2.1　マイクロホンアレー入力の周波数領域での分布

対象音源、および単一の雑音源から音響信号が平面波で到来する場合、対象音源と雑音源の方向が異なると、それらの信号はマイクロホンによって異なった時間差をもって加算されることになる。各マイクロホンで観測した信号にそれぞれ適当な遅延を加え対象音源からの信号の時間差を補正することにより i 番目のマイクロホンでの観測波形 $m_i(t)$ を

$$m_i(t) = s(t) + n(t - \tau_i) \tag{1}$$

と表せる。ここで $s(t)$ と $n(t)$ はそれぞれ対象音源と雑音源の時刻 t の信号、τ_i は i 番目のマイクロホンでの雑音信号の時間遅れである。これを共通の分析フレームでそれぞれ離散時間フーリエ変換すると周波数 ω の成分は

$$M_i(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau_i} \tag{2}$$

となる [1]。

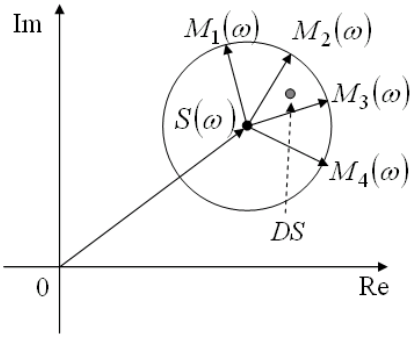

図 1: 観測信号の周波数成分の複素平面上での分布

式 (2) を幾何学的に考えると図 1 に示すように、$M_i(\omega)$ は複素平面上において全て $S(\omega)$ を中心とし半径 $\|N(\omega)\|$ の円上に分布する。

2.2　対象音源のスペクトルの推定

前節の議論より、3 個以上のマイクロホンについて $M_i(\omega)$ を観測し、複素平面上で多角形の外接円の中心を求めれば対象音源の短時間スペクトル $S(\omega)$ を求めることができる。これに対し、DS によって求まる信号の短時間スペクトルは図 1 において多角形の重心にあたるので真のスペクトルと一致しない。

それぞれの $M_i(\omega)$ の観測値には誤差が含まれていることを考慮して、本方法では $S(\omega)$ の推定値として

$$\tilde{S}(\omega) = \underset{Y(\omega)}{\operatorname{argmin}} \operatorname{Var}[\|\mathrm{M_i}(\omega) - \mathrm{Y}(\omega)\|^2] \tag{3}$$

で表される $\tilde{S}(\omega)$ を用いた。ただし $\operatorname{Var}[\cdot]$ は $\cdot$ の分散を表す。

多角形の外接円の中心を求める演算は雑音のスペクトルに対して非線型な処理であるので、線形フィルタによる処理とは性質が根本的に異なる。また、適応学習処理を必要とせず、フレーム単位で信号スペクトルが推定できる。

2.3　本方法の性質

(1) 雑音源が複数である場合の性質

実環境においては複数の雑音源がある場合がしばしばある。本方法では、雑音源は 1 つであることを仮定しているが、信号スペクトル推定はフレームごと周波数ごとに独立して行われるので、周波数ごとに雑音方向が別であっても構わないことになる。従って、雑音源が複数あっても、ある分析フレームの、ある周波数 ω について、ある雑音源が特に優勢であるならば、上述の原理は適用できることになり、雑音低減効果が期待できる。たとえば、雑音源が音声である場合、以上の性質が当てはまる可能性がある。

(2) 誤差に対する性質

式 (1) に対して式 (2) は近似表現であること、分析フレーム内で周波数 ω の雑音の方向が複数存在すること、マイクロホンの特性にばらつきがあること、

---

* “Speech recognition under noisy environment using microphone array and geometrical processing in frequency domain,”
by Takashi OKAJIMA, Yutaka KAMAMOTO, Takuya NISHIMOTO and Shigeki SAGAYAMA (The University of Tokyo).

[1] ただし、ここでは短時間フーリエ変換を考えており、異なる d_i に対して分析フレームは雑音信号の異なる部分を切り出すことになるので厳密には式 (2) の等号は必ずしも成り立たない。

表 1: 単語正解精度 (%) の比較 (シミュレーション)

マイク数	雑音源 1 個		雑音源 2 個		雑音源 3 個		雑音源 5 個	
	DS	本方法	DS	本方法	DS	本方法	DS	本方法
1	42.2		-5.7		-27.4		-27.5	
3	51.1	**87.1**	13.2	**57.0**	-12.2	**1.0**	-14.8	**-14.0**
4	54.1	**87.9**	10.5	**21.6**	-2.6	**20.1**	-8.1	**0.3**
8	55.1	**88.0**	22.9	**73.6**	7.7	**60.6**	8.2	**49.0**
16	55.7	**88.0**	25.3	**75.1**	11.2	**60.6**	18.6	**55.5**

表 2: 雑音源数及び対象音源からの角度

雑音源数	角度 (度)
1	30
2	30, 270
3	30, 60, 270
5	30, 60, 120, 180, 270

などの原因により、$M_i(\omega)$ の観測値は式 (2) の値と誤差を生じることが考えられる。特に、$M_i(\omega)$ の観測値が複素平面上で直線に近くなってしまうと、外接円の中心の推定は不安定な問題になってしまう。対象音源の短時間スペクトルを推定する方法をこれらの誤差に頑健になるように工夫するとさらに性能が向上する可能性がある。

本稿に挙げた実験においては、$M_i(\omega)$ の観測値の実数成分と複素数成分の相関係数の 2 乗が 0.99 を越えた場合には DS と同じように $M_i(\omega)$ の重心を推定値とした。

(3) マイクロホンの配置と周波数に対する性質

各マイクロホンの配置や、雑音の到来方向によってはある周波数 ω においてマイクロホン間の雑音の位相差が 2π の整数倍に近くなり、図 1 において $M_i(\omega)$ が円周上のせまい範囲に集中してしまい外接円の中心を精度良く求めることが難しくなる。マイクロホンの配置を工夫するとこのような状況が生じにくくなり、さらに性能が向上する可能性がある。

3 評価実験

3.1 実験の概要

本方法の評価のため、計算機上のシミュレーションおよび実環境で連続音声認識実験を行った。

実験は雑音を付加した音声について、本方法と DS での単語正解精度 (Word Accuracy) を比較した。

音声認識エンジンには Julius3.3p3 を用い、評価データには ASJ-JNAS の新聞記事読み上げ音声コーパスから男性、女性 100 文ずつを抜粋した IPA-testset を用いた。付加する雑音としては同一のコーパスから IPA-testset に含まれない 10 文を選び、用いた。音響特徴量は 12 次の MFCC とその Δ MFCC および Δ 対数 Power の計 25 次元とし、フレーム長 25ms・フレームシフト 10ms で分析した。

3.2 シミュレーション実験

シミュレーション実験ではマイクロホンアレーは直径 30cm の円周上に等間隔に 3 個から 16 個のマイクロホンを配置した。対象音源および雑音源はマイクロホンと同一の平面上にあるとし、雑音源の個数ならびに対象音源の方向に対する雑音源の方向の角度を表 2 のように変化させた。それぞれの雑音源のパワーは対象音源のパワーに対し −10dB とした。

表 3: 単語正解精度 (%) の比較 (実環境)

	マイク数 1	DS	本方法
雑音なし	37.4	**61.1**	59.5
雑音あり	9.3	34.1	**38.2**

結果を表 1 に示す。各表におけるマイク数 1 の条件はマイクロホンアレーを用いない場合の単語正解精度である。全ての条件において本方法では DS より認識性能が向上した。雑音源が単一の場合はクリーン音声での単語正解精度 89.4%に近い性能が得られ、雑音の影響をほぼ除くことができた。雑音源が複数の場合にも DS と比較して有効な方法であることが示された。

3.3 実環境での実験

実環境での実験ではマイクロホンアレーは 4 行 4 列の間隔 10cm の格子点状に 16 個のマイクロホンを配置した。マイクロホンを配置した平面に垂直な方向に対象音源を配置し、雑音源は対象音源から 30 度の方向に配置した。対象音源及び雑音源はマイクロホンアレーからほぼ 1m の距離に置いたスピーカを用いて再生した。雑音源のパワーは対象音源のパワーに対し −10dB とした。計算機などの発生する雑音がある通常の実験室において音声雑音を加えず対象音源のみ再生した場合と音声雑音を加えた場合について認識実験を行った。

結果を表 3 に示す。本方法における単語正解精度は DS と同程度に留まった。これは実験室の残響が比較的長く、その影響が大きかったためであると考えられる。

4 まとめ

マイクロホンアレーを用い、周波数領域での幾何学的な考察に基づいて、雑音環境中での対象信号の短時間スペクトル推定の方法を提案した。シミュレーションによる実験で DS と比較して認識性能を大きく向上させることができた。

今後の課題として、本方法を 2.3 節で挙げた性質についてより頑健にするための検討を行いたい。比較的残響の少ない環境において実験を行い本方法における残響の影響について検証し、本方法を改良したい。さらに、非定常な雑音源が時々刻々と移動する場合について本方法の適用可能性を実験によって検証したい。

参考文献

[1] 大賀寿郎, 山崎芳男, 金田豊:音響システムとディジタル処理, 電子情報通信学会, 1995.

[2] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄:音声認識システム, オーム社, 2001.