空間スペクトルへの周波数分析の適用による時間断続信号の検出*

☆杉本侑哉，加藤通朗，牧野昭二，山田武志（筑波大）

1　はじめに

会議や打合せを収録したアーカイブの効率的な再生のためには，誰が，いつ，どのように発話しているかという発話状況を自動的に検出し，タグ付けしておくことが有効である．このような効率再生の一環として，我々は音響信号のみを収録したアーカイブを対象とし，アーカイブ中の発話状況を判定する試みを行っている．

発話の意味・内容や意図を含む発話状況を判定するにあたっては，まず「連続信号」,「時間断続信号」,「無信号」という音響イベントを検出する必要があると考えられる．連続信号は，ある方向から時間的に連続して音響信号が到来している状況を指す．同様に時間断続信号は，ある方向から，相槌のように発話の時間長が短く，かつそれらが時間的に断続している音響信号が到来している状況を指す．これまでに我々は，このような音響イベントを空間スペクトルに対する主成分分析を用いて検出する手法を提案している[1].

本稿では，時間断続信号の検出に適用できる可能性がある新たな手法として，空間スペクトルに対する周波数分析について検討する．

2　提案手法

2.1 MUSIC による空間スペクトルの算出

会議用テーブルの中央にマイクロホンアレーを設置し，各マイクロホンで収録された音響信号を用いて MUSIC（MUltiple SIgnal Classification）法[2]により空間スペクトルを算出する．本稿では，MUSIC 法を周波数領域に拡張した手法[3]を用いる．

フレーム t ，周波数 ω におけるマイクロホンアレーへの入力音響信号ベクトルを

$$\mathrm{X}(\omega,t)=[\mathrm{X}_1(\omega,t),\cdots,\mathrm{X}_M(\omega,t)]^H \qquad (1)$$

と表す．ここで M はマイクロホン数である．このとき空間相関行列を次式により求める．

$$\mathrm{R}_{xx}(\omega,t)=E[\mathrm{X}(\omega,t)\mathrm{X}^H(\omega,t)] \qquad (2)$$

音源数を N とし（ただし $M>N$ とする），空間相関行列の小さい方から $M-N$ 個の固有値に対応する固有ベクトル $\mathbf{e}_i$ を用いて，空間スペクトルを以下のように算出する．

$$P(\omega,t,\theta)=\frac{1}{\sum_{i=N+1}^{M}\left|\mathbf{e}_i^H\mathbf{a}(\omega,\theta)\right|^2} \qquad (3)$$

ここで θ は方向，$\mathbf{a}$ は方向ベクトルを表す．本稿では，関心のある周波数領域 ω_j ～ ω_k において平均化した空間スペクトル $P(t,\theta)$ を用いる．

2.2 空間スペクトルに対する周波数分析

MUSIC により求めた空間スペクトルから，方向ごとに連続する L フレームを抽出し，離散フーリエ変換を適用する．なお，周波数スペクトルに強い特徴を与えるために，(0 , 1) に二値化した空間スペクトルを用いる．このとき，

$$\left|Y(k,\theta)\right|=\left|\sum_{t=1}^{L}P(t,\theta)e^{-j\frac{2\pi nk}{L}}\right|,k=1,2,\cdots,L \qquad (4)$$

である．ここで，$|Y(k,\theta)|$ は方向ごとの振幅スペクトルである．これは，各方向の音響信号のパワーの時間変化を周波数分析していることに相当する．提案手法では，このようにして求めた振幅スペクトルの形状から，連続信号，時間断続信号，および無信号のいずれであるかを判別する．

3　提案手法の有効性の評価

提案手法の有効性を検証するために，インタビュー形式の複数人の会話を収録した．収録音響データには，発話に相当する連続信号，および相槌に相当する時間断続信号が多く含まれている．

収録は，Fig.1 に示すように，セミナー室の中央付近のテーブル上に設置したマイクロホンアレーを用いた．マイクの数は 8 であり，各マイクは直径 20 cm，高さ 10 cm の台の円

*Scattered Signal Detection by Frequency Analysis of Spatial Spectrum, by Yuya SUGIMOTO, Michiaki KATOH, Shoji MAKINO, and Takeshi YAMADA (University of Tsukuba).

周上に均等に設置されている．テーブルのサイズは 1.8 m×1.2 m である．発話者は 3 名であり，うち 1 名はインタビュアーである．マイクロホンアレーから各話者までの距離は概ね 1.0 m である．このような状況において 15 分程度の収録を行った．なお，サンプリング周波数は 16 kHz である．

収録した音響信号に対して，1 フレームを 0.5 秒として，Fig.2 上段のような 5° 刻み（72 方向）の空間スペクトルを求めた．なお，MUSIC におけるフレーム長は 512 ポイント，フレームシフトは 128 ポイントであり，0.5 秒の時間平均を求めている．また，MUSIC における周波数平均の範囲は 1～4 kHz である．

このようにして得られた空間スペクトルに対し，方向ごとに 15 秒（L=30）のフレーム群を取り出し，フレーム長 1024 ポイントで離散フーリエ変換を適用する．

Fig.2 は，空間スペクトルに対する離散フーリエ変換の結果の例である．本稿では，連続した 120 フレームを用いて，180° 方向の話者に対して分析を行った．上段は分析区間の空間スペクトルである．横軸はフレームを表し，縦軸は方向を表している．中段は，上段の空間スペクトルに対して，0.5 を閾値にとって二値化した図である．下段は，二値化した空間スペクトルに離散フーリエ変換を適用して得られたスペクトログラムである．横軸はフレームを表し，縦軸は周波数を表している．連続信号（図中の①）の場合は，低周波成分のパワーに強い特徴が現れている．一方，時間断続信号（図中の②）の場合は，連続信号よりも低周波成分のパワーが弱く，広い周波数域にパワーが存在していることが読み取れる．このような傾向に基づいて連続信号と時間断続信号との判別が可能になると考えられる．

4　おわりに

会議中に発せられる相槌のような時間断続信号について，空間スペクトルの周波数分析による検出が可能かどうかを検討した．

連続信号との比較により，時間断続信号は特に高周波成分を多く含むことが確認されたため，周波数分析後の周波数スペクトルのパターンマッチングを用いて両者の判別が可能であることが示唆された．

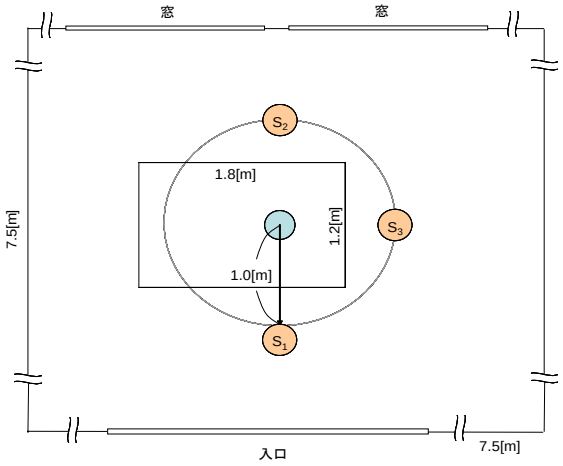

Fig. 1 使用する音声データの収録環境

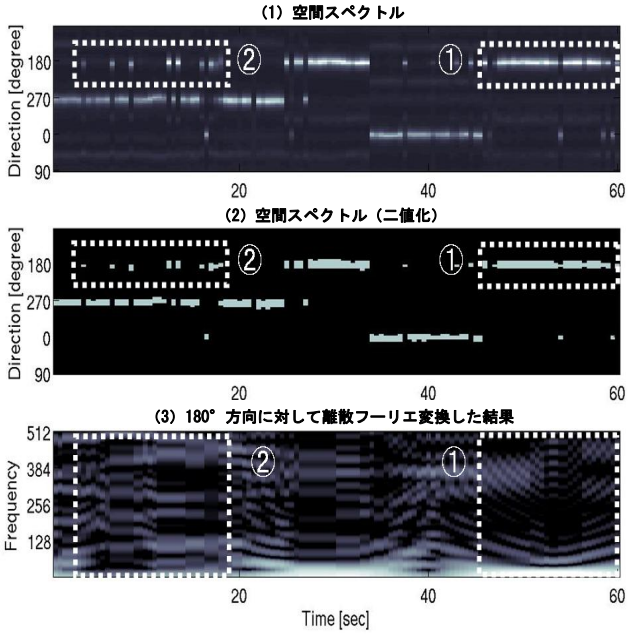

Fig. 2 連続信号に対する FFT の結果

謝辞

本研究についてのご助言をいただいた，東北大学大学院工学研究科・伊藤彰則教授に謝意を表します．

参考文献

[1] 加藤他, 信学技法, EA2010-47, pp.25-30, 2010.

[2] Schmidt *et al*, IEEE Trans. On antennas and propagation, Vol.AP-34, No.3, pp.276-280, 1986.

[3] Asano *et al*, EURASIP journal on applied signal processing, pp.1727-1738, 2004.