空間的サブトラクションアレーに基づくハンズフリー音声認識システムの開発*

猿渡洋[1]，庄境誠[2]，長濱克昌[2]，山田真士[2]，西浦敬信[3]，傳田遊亀[3]，高橋祐[1]，鹿野清宏[1]
奈良先端大[1]，旭化成（株）新事業本部[2]，立命館大学[3]

1 はじめに

従来の音声対話システムは，環境雑音の影響を低減するために接話型指向性マイクを用いているが，ユーザの位置や姿勢が大きく制約を受けるといった問題があった．そこで著者らは，本 e-Society プロジェクトにおいて，ハンズフリーで音声認識が出来るシステムの開発を行った．具体的には，高精度な雑音抑圧手法として，空間的サブトラクションアレー (SSA) を提案した [1]．SSA は，死角制御型ビームフォーマ (NBF) により推定された雑音を，遅延和法 (DS) にて得られた目的音強調信号からスペクトル減算することにより，目的音を頑健に抽出する手法である．本稿では，SSA の原理・性能評価結果を述べると同時に，SSA を DSP 上にてリアルタイム実装して音声対話システムに組み込んだ例を紹介する．

2 空間的サブトラクションアレーの概要

2.1 アルゴリズム

SSA は Fig. 1 に示すように，DS を用いてユーザ方位 θ を同位相化し，雑音が多少残留したユーザ音声スペクトル $Y(k,\tau)$ を推定する主パスと，NBF を用いて ユーザ方位 θ に死角を作り，雑音スペクトル $Z(k,\tau)$ を推定する参照パスから成る．ただし，k は周波数番号 (ビン)，τ はフレーム番号である．そして，主パスから参照パスをパワースペクトル上で減算することで雑音抑圧を実現する．ただし，SSA は時間波形ではなく，音声認識を行う際の特徴量である Mel Frequency Cepstrum Coefficient (MFCC) [2] を出力するため，両パスに対して，メルフィルタバンク分析を行い，以下の式で減算する．

$$m(l)=\sum_{k=k_{lo}(l)}^{k_{hi}(l)} W(k;l)\{|Y(k,\tau)|^2-\alpha(l)\cdot\beta\cdot|Z(k,\tau)|^2\}^{\frac{1}{2}}$$
$$(\text{ if } |Y(k,\tau)|^2-\alpha(l)\cdot\beta\cdot|Z(k,\tau)|^2\geq 0\ ),\quad (1)$$
$$m(l)=\sum_{k=k_{lo}(l)}^{k_{hi}(l)} W(k;l)\{\gamma\cdot|Y(k,\tau)|\}\quad (\text{ otherwise }).\quad (2)$$

ここで，$W(k;l)(l=1,\cdots,L)$ はメルフィルタ分析窓であり，$k_{hi}(l)$ および $k_{lo}(l)$ はそれぞれ l 番目のフィルタの下限，上限周波数番号を示し，$m(l)$ はメルフィルタバンク上で雑音抑圧処理を行うことによって得られた l 番目帯域の振幅スペクトル和である．更に，β は各帯域に対して一定の減算係数であり，γ も各帯域に対して一定のフロアリング係数である．$\alpha(l)$ は，雑音のみが存在する区間で減算結果が 0 になるように各帯域で調整されるパラメータである．SSA で出力される MFCC は，式 (1)(2) で得られた $m(l)$ の対数値を，離散コサイン変換することで求められる [2]．

2.2 SSA の特徴

SSA は音声認識に特化したアレー信号処理であり，音声認識を行う際の特徴量である MFCC 係数を直接出力し，波形再合成等の冗長な処理を行わない．一般的に，音声認識では位相情報を必要としないため，パワースペクトル上で減算を行う SSA は効率的な信号処理とみなすことができる．また，パワースペクトル上でのスペクトル減算処理に基づいて雑音抑圧を行うため，多少の雑音推定誤差に対しても頑健であることが知られている．更に，調節するパラメータも，通常 24 個からなる $m(l)$ に対するパラメータ $\alpha(l)$ だけであり，高速に処理できる．

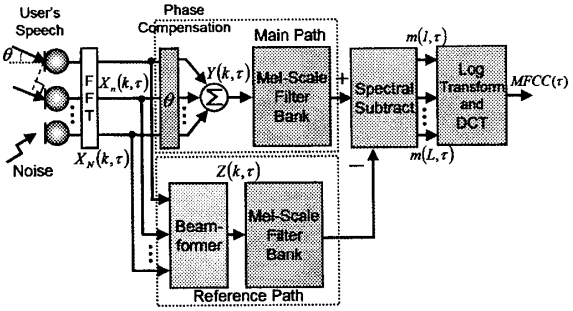

Fig. 1 SSA における信号処理の流れ．

3 実験

3.1 実験条件

提案法の有効性を確認するため，実環境を模擬した実験を行った．ここでは，4 素子または 8 素子のマイクロホンアレーを用いた．比較対照法として，シングルチャネルマイク入力 (信号処理無し)，DS，典型的な適応型アレーであるグリフィス・ジム型アレー (GJ) も同様に実験を行った．ここで，DS と SSA はほぼ同じ程度の演算量であるが，GJ は SSA に比べて多大な演算量がかかることに注意が必要である．

入力信号には，JNAS クリーン音声データベース [2] に Fig. 2 に示される部屋で計測したインパルス応答を畳み込み，音声 SNR が平均 5 dB となるように掃除機雑音を重畳したものを用いた．認識タスクは，新聞記事読み上げ (2 万語) のディクテーションである (認識実験条件の詳細は参考文献 [1] を参照のこと)．音声認識における音響モデルは，PTM [3](2000 状態，64 混合) のクリーンモデルに既知雑音をマッチドさせたものを用いた [1]．性能評価値として，単語認識精度 [2] を採用した．β，γ については，それぞれ音声認識性能を基に最適なものを選んだ．

3.2 実験結果および考察

Figure 3 に各手法の単語認識精度を示す．素子数が 4，8 共に，提案法が全ての従来法を上回っていることが分かる．特に，提案法は，素子数の少ない場合 (4 素子の場合) において，性能改善が顕著であった．SSA の演算量は GJ に比べて非常に少ないことを考慮すると，提案法がより現実的な雑音抑圧処理であることが伺える．

---

* "Development of hands-free speech recognition system based on spatial subtraction array," by Hiroshi Saruwatari[1], Makoto Shozakai[2], Katsumasa Nagahama[2], Masashi Yamada[2], Takanobu Nishiura[3], Yuuki Denda[3], Yu Takahashi[1], Kiyohiro Shikano[1] (NAIST[1], Asahi Kasei Corporation[2], Ritsumeikan Univ.[3]).

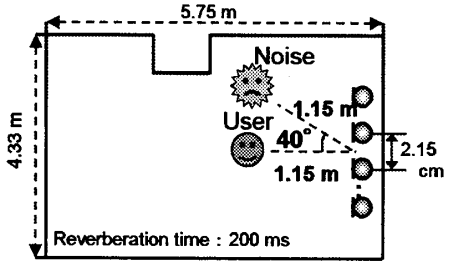

Fig. 2　実験に用いた残響室の概要.

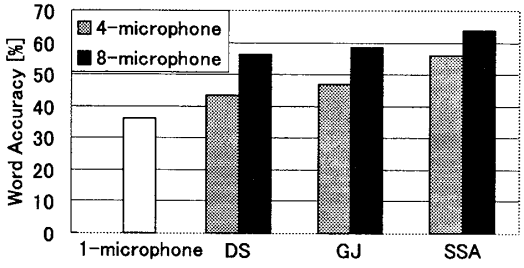

Fig. 3　各手法における音声認識結果.

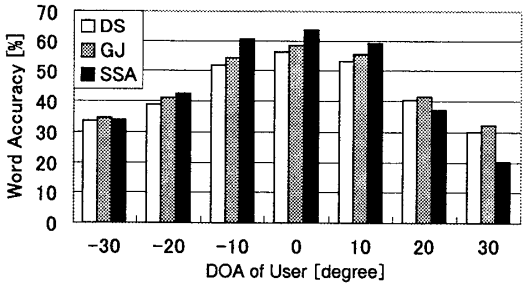

Fig. 4　ユーザ方位がずれた場合の音声認識結果.

次に，ユーザが正面方位からずれた場合の音声認識精度を Fig. 4 に示す (ここでは 8 素子の場合のみ). 提案法は，±20° 程度までは，従来法とほぼ同等かそれ以上の耐誤差性能を有していることが分かる.

4　DSP 版ハンズフリー音声認識システム

提案する SSA をハードウェア上にて実時間動作させるために，DSP 版のハンズフリー音声認識システムを開発した. DSP ボードの具体的な仕様を Table 1 に，システムの概観を Fig. 5 に示す.

本 DSP ボード内部には，(1) 多チャネル AD/DA 変換器，(2) 話者方位推定に基づくリアルタイム発話検出器 [4]，(3) リアルタイム動作可能な SSA アルゴリズム，(4) MFCC パラメータを TCP/IP にて伝送するネットワークモジュール，等が実装されている. 本ボードを直接ネットワーク経由で音声認識装置へ接続することにより，ハンズフリー音声認識システムを容易に構成することができる. 現在，本プロジェクトで開発された音声対話システム「たけまる君」に接続することにより，遠隔発話にも対応した音声案内システムが実現されている.

5　まとめ

本稿では，高精度な雑音抑圧を可能とする空間的サブトラクションアレーについて解説を行った. また，提案アルゴリズムを DSP ボード上に実時間実装した装置も紹介し，ハンズフリー音声認識システムの構成例を示した. 本 e-Society プロジェクトで開発された SSA は，音声認識に特化した特徴量をリアルタイムで出力できることから，容易に従来の音声対話システムと結合することが出来るのが特徴である. 今後は，様々なハンズフリー音声処理システムにて，本提案システムが活用されることを期待するものである.

Table 1　DSP ボードの仕様

ボード	MTT 社製 DSP ボード (TMS320C6713) MTT 社製 Ethernet ボード
入力	アナログ 8 チャネル
出力	アナログ 2 系統，ディジタル 1 系統
AD 変換	旭化成社製 AKM AK5384 (8, 16, 24, 48 kHz 可変)・24 bit 精度
DA 変換	旭化成社製 AKM AK4380 48 kHz・24 bit 精度
電源	5 V

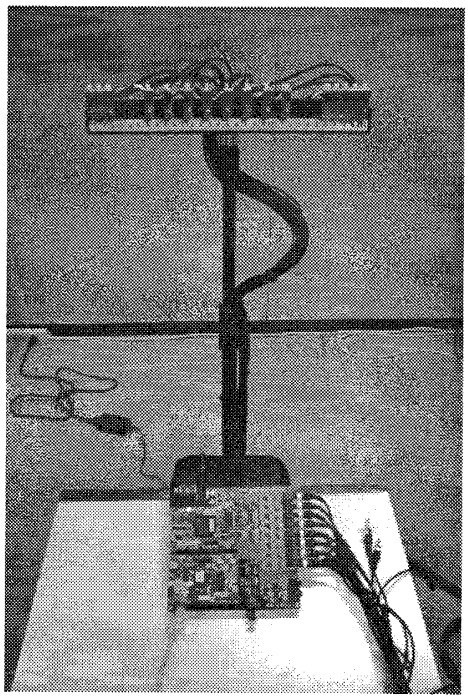

Fig. 5　DSP 版ハンズフリー音声認識システム.

謝辞　この研究の一部は，文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われたものである.

参考文献

[1] Y. Ohashi, et al., *Proc. of IROS2005*, pp.533–537, 2005.
[2] 鹿野 他, 音声認識システム, オーム社, 2001.
[3] A. Lee, et al., *Proc. of ICASSP*, vol.III, pp.1239–1272, 2000.
[4] 西浦 他, " 話者方位推定に基づくリアルタイム話者区間検出システムの開発," 情報処理学会第 70 回全国大会, 2008 年 3 月.