24ｂｉｔオーディオボードを使ったワイドレンジ耐雑音音声認識

Wide-Range Noise Robust Speech Recognition using a 24-bit Audio Board

○荒川 隆行、田中 大介、西沢 俊弘、山下 信行
(ＮＥＣ共通基盤ソフトウェア研究所)

* Takayuki ARAKAWA, Daisuke TANAKA, Toshihiro NISHIZAWA, Nobuyuki YAMASHITA,
(NEC Common Platform Software Research Laboratories)

t-arakawa@cp.jp.nec.com, d-tanaka@rf.jp.nec.com, nishizawa@bk.jp.nec.com, n-yamashita@ax.jp.nec.com

Abstract— For some voice input systems; such as robots or car-navigation systems, distances between speakers and microphones are variable. In these cases, SNR often becomes worse and the dynamic range of voice volumes becomes wide. To deal with these noisy and wide-range voices, this paper proposes a combined method with a 24-bit audio board and noise suppressor. Speech recognition experiments shows that the performance with 24-bit audio recording is better than 16-bit audio recording and that, for noisy environment, 24-bit audio recording with noise suppressor achieves much better performance than without it.
Key Words: Speech Recognition, Gain Control, Robot.

1　はじめに

近年、ロボットやカーナビ等の様々な機器のインターフェースとして、音声対話機能が注目されている。しかしながら、これらの機器は離れた位置から音声コマンドで制御を行うため、ヘッドセットを使った近接発話に較べると性能が下がってしまうという問題がある。

この性能劣化の主な原因として、周囲雑音の影響とマイクゲインの不一致という２つの原因が考えられる。前者の周囲雑音の影響は、マイクと発話者との距離が離れるため周囲の雑音の影響を受け易くなりＳＮＲが低くなる事に起因する。後者のマイクゲインの不一致は、話者毎の発声音量の違いや発話者とマイクとの距離が一定でないために音量のレンジが広くなることに起因する。マイクゲインが合っていないと、音割れや量子化誤差による性能劣化が起こる。

前者の周囲雑音の影響を軽減する方法としては、雑音成分を推定し抑圧するノイズサプレッサ[1]や、複数マイクを用いて対象となる音声を強調するマイクロホンアレー[2]など様々な方法が提案されている。

後者のマイクゲインの不一致に対しては、マイクゲインを動的に変更するAutomatic Gain Control (AGC)が用いられてきた[3][4]。しかしながら、動的にマイクゲインを変更することは、特徴量の差分成分に悪影響を及ぼすなど音声認識の性能劣化につながることが知られている。

本稿では、上記周囲雑音の影響とマイクゲインの不一致という２つの性能劣化の原因に対し、２マイクノイズキャンセラと、24ビットオーディオボードを組み合わせたワイドレンジ耐雑音音声認識を提案する。本稿では、まず、２節で提案法の全体の構成について説明する。次に、３節で複数チャンネルのマイクに対して24ビットのオーディオデータを収録できるオーディオボードについて説明する。次に、４節で対象話者方向以外から来る妨害音を除去するノイズキャンセラについて説明する。次に、５節で24ビットのオーディオデータを音声認識向けに16ビットに変換する方法について説明する。次に６節で、今回用いた音声認識について説明し、最後に７節で複数の音量、複数のマイク距離で収録した音声データに対する音声認識の評価について説明する。

2　全体構成

全体の構成をFigure.1に示す。雑音のある環境で対象話者の音声を強調し認識するために、以下のような手順で処理を行う。

1. まず、音声マイクと雑音マイクを用いて、対象話者方向の音声と、対象話者方向以外から来る妨害音を取得する。
2. 次に、24ビットオーディオボードでこれら２つのマイクから取得されたアナログデータを24ビットのデジタルデータに変換する。
3. 次に、ノイズキャンセラで、この２つのデジタルデータから対象となる音声のみを強調し取り出す。
4. 次に、24→16ビット変換で、音声認識の入力に合うように24ビットのデジタルデータを16ビットに変換する。
5. 最後に音声認識を行う。

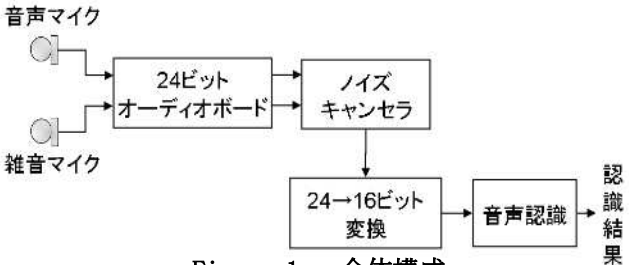

Figure 1.　全体構成

3　24 ビットオーディオボード

通常、マイクより取得されたアナログ音声データは、16ビットのA／D変換器によりデジタルデータに変換される。しかしながら、機器から数十ｃｍに近づいて入力される音声と、機器から数m離れたところから入力される音声とでは、音量のダイナミックレンジが異なるため、16ビットの範囲に収まらない。そこでよりレンジの広い音声を収録できる24ビットオーディオボード(DS-BD-24ADUSB)を試作した。今回開発したオーディオボードのスペックをTable. １に示す。

Table.1 24ビットオーディオボードのスペック

サンプリング周波数	48kHz, 32kHz, 16kHz, 8kHz, 44.1kHz, 22.05kHz, 11.025kHz（本稿では11kHzを使用）
チャンネル数	8 ch（本稿では2chのみ使用）
分解能	24 bit
サイズ	W100mm x D60mm x H20mm
インターフェース	USB 2.0
ノイズ性能	105dB

ノイズ性能とは、マイクを接続しない状態で計測したノイズレベルと最大入力レベルとの比を意味する。24ビットA／D変換の理論上の最大値は120dBである。参考までにほぼ同等の構成の16ビットオーディオボードでのノイズ性能は70dBである[5]。今回試作したボードは、マイクゲインを変更することなく35dB広いレンジの音声を扱うことができる。

4　ノイズキャンセラ

ノイズキャンセラは、音声マイクに混入する雑音成分を雑音マイクを用いて推定し、消去することで対象とする音声のみを強調する。一般的なノイズキャンセラの構成を Fig. 2 に示す。音声マイクに混入する雑音成分を推定するために、雑音マイクに入力された信号に対し適応フィルタの処理を行う。音声が大きいときは雑音の推定にとって音声は妨害信号となるため、適応フィルタのステップサイズを小さくし、更新を抑える。音声が小さく雑音が大きいときは雑音への追従性能を高めるため、適応フィルタのステップサイズを大きくする。

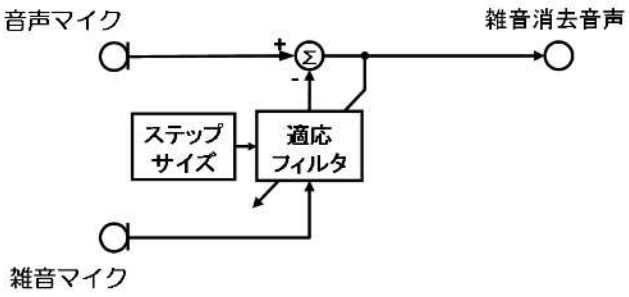

Figure 2.　一般的なノイズキャンセラの構成

しかしながら、このような一般的なノイズキャンセラでは、雑音マイクに混入する音声信号（クロストーク）を誤って雑音としてしまうために、対象とする音声を消去してしまう。

本稿では、このようなクロストークの影響を軽減するために、クロストークの推定、消去を行うノイズキャンセラを用いた。Fig. 3 にノイズキャンセラの構成を、Fig. 4 に推定 SN 比と係数更新ステップについて示す。提案法のノイズキャンセラは、クロストーク推定用の適応フィルタも備えている。雑音用フィルタは音声用マイクロホンから雑音を消去し、音声用フィルタは雑音用マイクロホンに混入する音声を消去する。上段にある２つのフィルタがメインフィルタ、下段にある２つのフィルタがサブフィルタである。雑音成分用と音声成分用の適応フィルタは、それぞれ専用のサブフィルタを用いて推定した SN 比でステップサイズを制御する。SN 比が大きく音声が支配的なときは音声用フィルタのステップサイズを大きくし、雑音用フィルタのステップサイズを大きくする。反対に SN 比が小さく座右音が支配的なときは雑音用フィルタのステップサイズを大きくし、音声用フィルタのステップサイズを小さくする。このような構成とすることにより、大きな消去量と小さな音声歪を両立する[6]。

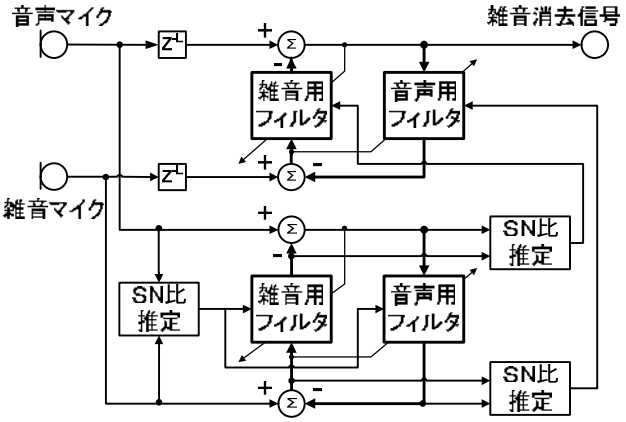

Figure 3.　提案法のノイズキャンセラの構成

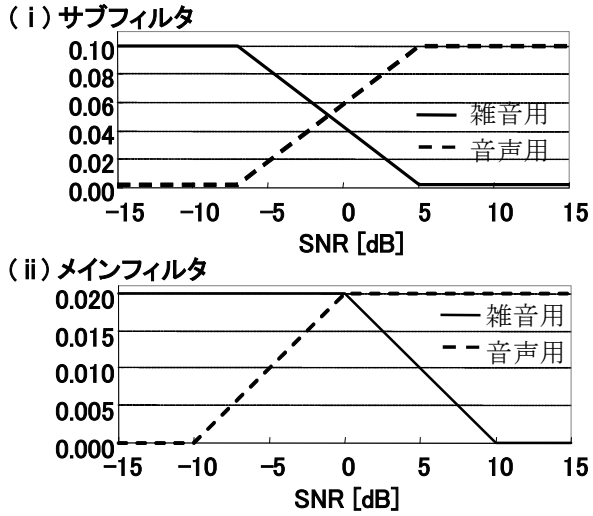

Figure 4.　推定SN比と係数更新ステップサイズ

5　24 ビットから 16 ビットへの変換

一般的な音声認識システムは、入力が16ビットであるため、本稿では16ビットのオーディオデータを入力とする音声認識システムを用いた。この為、24ビットのオーディオデータを16ビットに変換し、音声認識システムへの入力とした。変換には以下の３つの方法を実装した。

- **下位８ビットを削る**
  24ビットのオーディオデータに対し、下位８ビットを削り、上位16ビットのみを用いる。この場合は最大入力レベルを揃え16ビットで音声を収録するのと同等である。下位ビットを削ってしまうために、音量の小さな音声は量子化誤差の影響を受けやすくなる。

- **非線形処理を行い、ビット数を削減**
  Figure 5 に示す非線形関数を用いて、入力値に対して出力値を計算する。図の横軸が入力値（振幅の値）、縦軸が出力値である。24ビットのデータから符号ビットを除いた下位４ビットを削り、上位５ビット分に対し音量抑圧（コンプレッサ）を行った。この方法では、前記下位８ビットを削る方法に較べ扱える音量のレンジは広くなるが、音量の大きな音声が歪んでしまう。

- **適切な16ビットを選択する**
  環境と話者毎に予め音量の最大値を求め、その最大値が16ビットの最大値に収まるように適切な16ビットを選択する。この場合はマイクゲインを理想的に正しく設定した16ビットでの音声収録と同等となる。

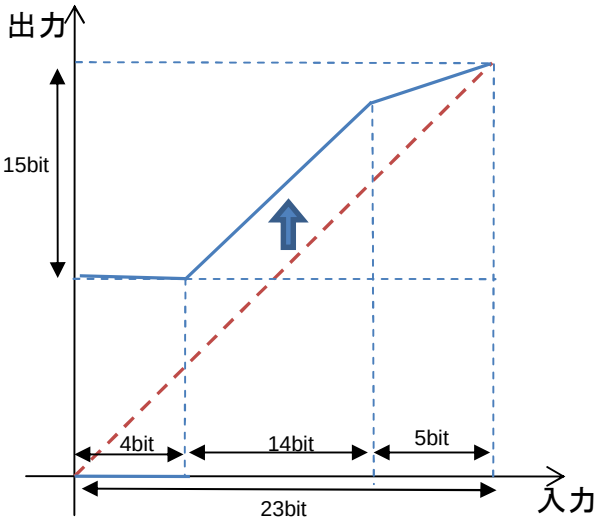

Figure 5. 非線形関数

入力は符号ビットを除く 23 ビット、出力は符号ビットを除く 15 ビットである

6　音声認識

音声認識には、単語認識エンジンを用いた。音声認識辞書は，W3C 勧告済の記述仕様である SRGS (Speech Recognition Grammar Specification)[7]のサブセットに対応している[8]。本稿では５０単語の辞書を用いた。特徴量にはケプストラムの１次から 10 次までの成分、およびその差分成分、パワーの差分、調波性特徴量およびその差分成分の計２３次元の特徴量を用いている。

7　評価実験

7.1 音量の異なる音声の収録

発声音量およびマイクとの距離の異なる音声をスピーカーより再生し、24ビットで収録した。以下の評価においてマイクゲインは一定である。スピーカー及びマイクの配置をFig. 6に、収録風景をFig.7に示す。

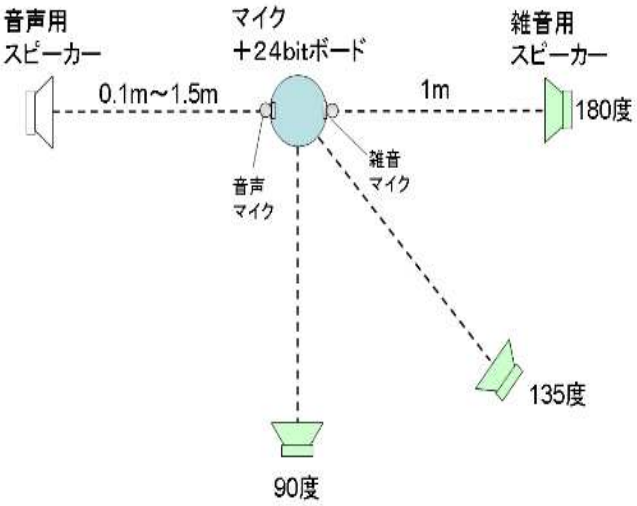

Figure 6. スピーカー、マイク配置

Figure 7. 撮影風景

7.2 音量の変化に対する評価

まず、音量の変化に対する24ビットでの音声収録単体の効果を見るために、雑音が無い環境で、ノイズキャンセラをオフにした評価を行った。評価条件をTable 2に示す。評価には音声認識を用い、単語正解率を求めた。

Table 2. 評価条件

話者	男性2名、女性2名、子供3名 （スピーカー再生）
発声内容	50単語
発声音圧※	50dBA，60dBA，70dBA
音声マイクとの距離	0.1m，0.5m，1.0m，1.5m

※発声音圧は、スピーカーとマイクを1.0m離した状態で、音声マイクの位置で計測した。

- **下位 8 ビットを削った場合の評価**

収録した音声に対し、まず下位8ビットを削った場合の評価を行った。結果をFig. 8に示す。50dBAの発声ではマイクとの距離が長くなるにつれて大きく性能が劣化している。これは音量が小さくなり16ビットでは量子化誤差が大きくなった為と考えられる。また、70dBAの発声においてマイクとの距離が0.1mの時に性能が悪くなっている。先程とは逆に音声が大きくなりすぎたために音割れが起きた為と考えられる。

- **非線形処理を行った場合の評価**

次に、非線形処理を行ってビット数を削減した場合について評価を行った。結果をFig.9に示す。50dBAの音声の劣化がなくなっていることがわかる。しかし、70dBAの音量が大きい音声の認識性能が若干劣化している。これは、非線形処理の影響により音量の大きい成分に歪みが生じたためであると考えられる。

- **適切な 16 ビットを選択した場合の評価**

次に、適切な16ビットを選択した場合の評価を行った。結果をFig.10に示す。上記２つ(Fig.8, Fig.9)に較べ音声認識率の劣化がなく最も性能が高くなっていることがわかる。

以上の評価から、24ビットで音声収録を行うことで、従来の16ビットの音声収録では対応できなかった広いレンジの音声に対してマイクゲインの変更無しに対応できることが確認できた。

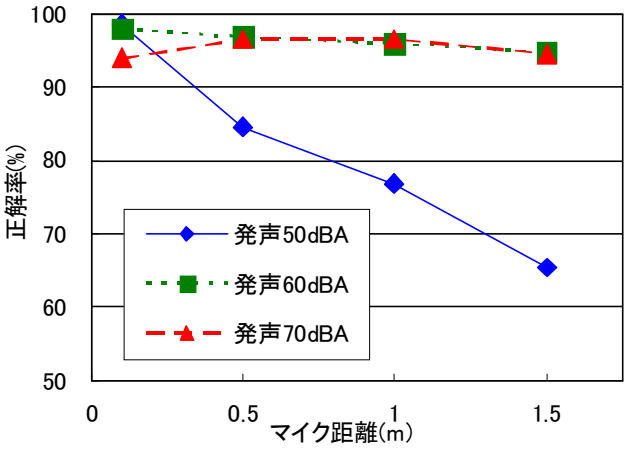

Figure 8. 下位８ビットを削った場合

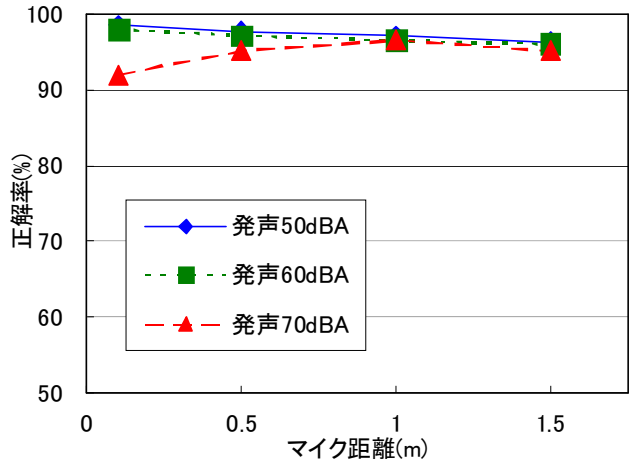

Figure 9. 非線形処理を行った場合

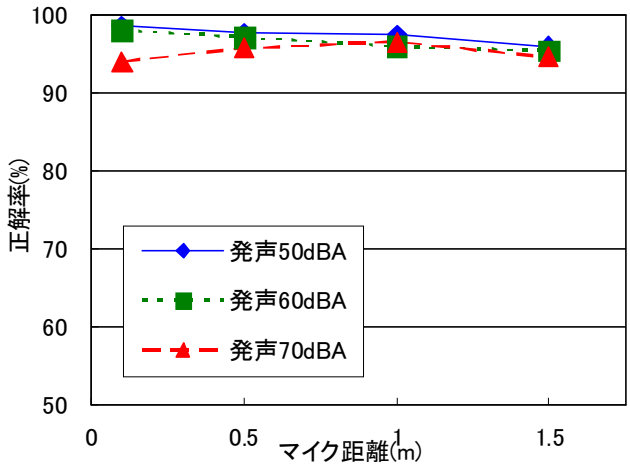

Figure 10. 適切な16ビットを選択した場合

7.3 ノイズキャンセラの効果

次に、ノイズキャンセラの効果を見るため、ノイズキャンセラを有効にし、雑音のある環境と雑音の無い環境での評価を行った。評価条件をTable 3に示す。以下の評価では、対象となる発声の音圧70dBAとし、雑音の音圧を60dBAに固定している。対象となる発声とマイクとの距離および雑音方向を変化させて評価を行った。先程と同様に評価には音声認識を用い、単語正解率を求めた。

Table 3. 評価条件

話者	男性2名、女性2名、子供3名 （スピーカー再生）
発声内容	50単語
発声音圧※	70dBA
音声マイクとの距離	0.1m，1.0m，1.5m
雑音内容	テレビCM（スピーカー再生）
雑音音圧※	60dBA
雑音マイクとの距離	1m
雑音方向	180度、135度、90度 （対象発話方向を0度とする）

※発声音圧および雑音音圧は、スピーカーとマイクを1.0m離した状態で、音声マイクの位置で計測した。

- **下位 8 ビットを削った場合の評価**

下位8ビットを削った場合の評価結果をFig. 11に示す。発声用スピーカーと音声マイクとの距離を1.5m, 1.0m, 0.1mと変えたものをプロットしている。図中の中塗りの印はノイズキャンセラを行った結果を示し、白抜きの印はノイズキャンセラを行っていないことを示す。雑音の有る環境では、３つの方向全てに対してノイズキャンセラの効果がはっきりと現れている。しかしながら、雑音が無い環境では、マイク距離1.5mの条件でノイズキャンセラを行うと性能劣化が見られる。これは、雑音マイクに回り込んだ音声を誤って雑音と判定してしまったためであると考えられる。

- **非線形処理を行った場合の評価**

非線形処理を行ってビット数を削減した場合の評価結果をFig.12に示す。雑音が有る環境では全ての条件に対しノイズキャンセラの効果が現れている。

- **適切な16ビットを選択した場合の評価**

適切な16ビットを選択した場合の評価結果をFig.13に示す。上２つ(Fig.11, Fig.12)と同様、雑音が有る環境で全ての条件に対しノイズキャンセラの効果が現れている。この場合が最も性能が高くなっている。

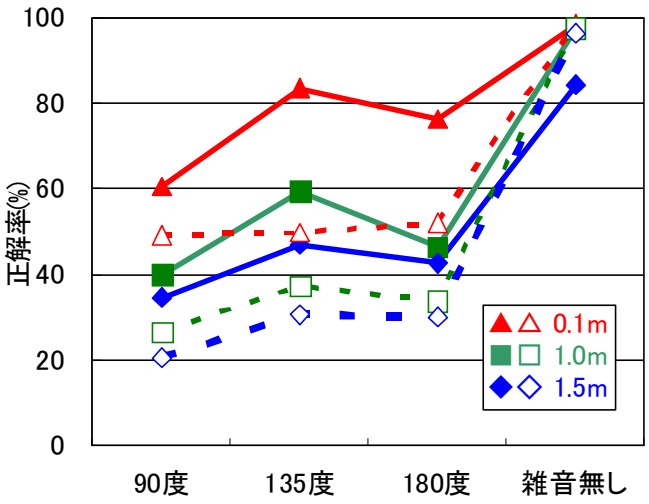

Figure 11. 下位８ビットを削った場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

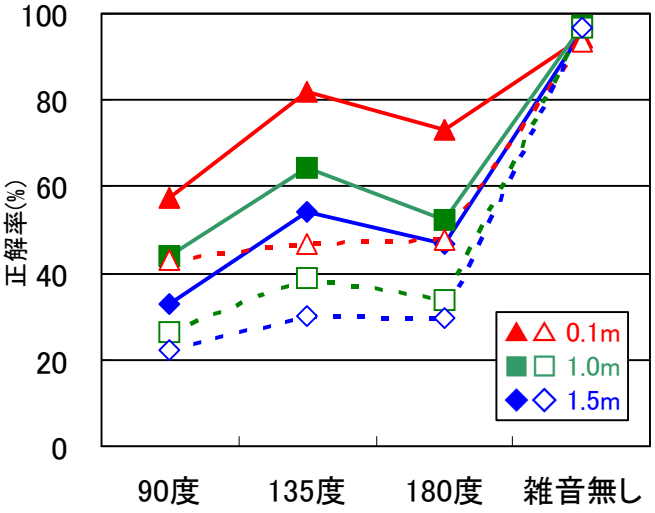

Figure 12. 非線形処理を行った場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

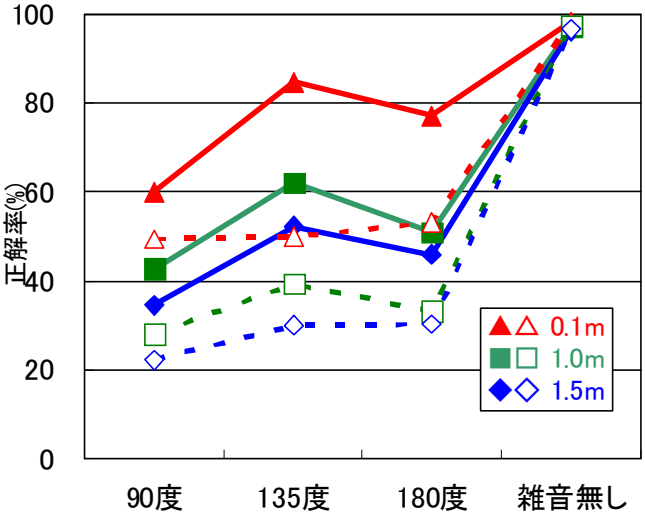

Figure 13. 適切な16ビットを選択した場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

8　まとめ

離れたマイクを用いて音声認識を行う際に問題となる、雑音とマイクゲインの不一致に対して、２マイクノイズキャンセラと24ビットオーディオボードを用いたワイドレンジ耐雑音音声認識を提案し、評価を行った。

24ビットで音声を収録することで、16ビットでは対応できなかった広いレンジの音声に対応できることを確認した。また、ノイズキャンセラを用いた評価では、雑音環境ではどの条件であってもノイズキャンセラを用いなかった場合に較べて大幅な性能改善が見られた。さらに適切な16ビットを選択した場合には、雑音の無い環境でノイズキャンセラを用いても性能劣化の無いことが確認できた。

9　今後の予定

本稿では、16 ビット入力に対応する音声認識での評価を行ったが、今後は 24 ビット入力に対応する音声認識で評価を行う予定である。今回評価した適切な 16 ビットを選択する処理はバッチ処理であり一発話分処理が遅延してしまう。24 ビット入力に対応した音声認識を用いることで 24 ビットから 16 ビットに変換する必要がなくなるために、遅延の無いオンラインでの処理が可能となる。

また、話者方向推定や音声検出と組み合わせるなど、さらに雑音環境での性能向上を行う予定である。

謝辞

本研究は、平成 20, 21 年 NEDO 委託研究『次世代ロボット知能化技術開発プロジェクト』の一環として行った。本研究をご支援いただいた関係各位に感謝する。

参考文献

1) M. Kato, A. Sugiyama and M. Serizawa, “A Family of 3GPP-standard Noise Suppressors for the AMR Codec and the Evaluation Results,” ICASSP ’03 SP-P5.14, 2003.
2) M. Brandstein and D. Ward, "Microphone Arrays," Springer Verlag, Berlin, 2001.
3) 小林他, “信学論”, J87A(12), 1491-1501, 2004.
4) 寺澤, 竹山, “マルチゲイン関数自動選択型音声 AGC”, 松下電工技法(Feb), 70-74, 2003.
5) 東京エレクトロンデバイス, “16 チャンネル専用 A/D・D/A ボード”, http://www.inrevium.jp/pm/image_audio/16adusb.html
6) M. Sato, A. Sugiyama, S. Ohnaka, "An Adaptive Noise Canceller with Low Signal-Distortion Based on Variable Stepsize Subfilters for Human-Robot Communication", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2055-2061, Aug. 2005.
7) http://www.w3.org/TR/speech-grammar/
8) 岩沢, ”組込み機器への搭載を可能にする小型音声対話モジュールの開発,” 機械設計, 第 51 巻, 第 17 号, pp.114-118, (2007).