社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（６）
2007／12／20

[招待講演] 新しい音声メディアによるユニバーサルコミュニケーション

鹿野 清宏†

†奈良先端科学技術大学院大学 情報科学研究科 〒630-0192 生駒市高山町 8916-5
E-mail: † shikano@is.naist.jp

あらまし 新しく発見した静かな音声メディアである非可聴つぶやき(NAM)と、新しい音源分離技術の SIMO-ICA をによるユニバーサルコミュニケーションへの挑戦について紹介する。とくに、声を出さずに電話できる無音声電話、発話障害者の発話補助、騒音下での明瞭なハンズフリー通話、両耳補聴器などの可能性について述べる。さらに、人と機械との音声対話システムとして、生駒市北コミュニティの音声情報案内システム「たけまるくん」、学研北生駒駅の「キタロボ」と「キタちゃん」、さらに、ハンズフリーロボット対話実験システムについても紹介する。
キーワード 音声メディア、ユニバーサルコミュニケーション、非可聴つぶやき（NAM），
ブラインド音源分離（BSS)、音声対話システム、音声情報案内システム、ロボット対話

[Invited Lecture] New Speech Media Applied to Universal Communication

Kiyohiro SHIKANO†

† Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan
E-mail: † shikano@is.naist.jp

Abstract We found a new quiet speech media, Non-Audible Murmur (NAM), and a new blind source separation algorithm, SIMO-ICA. These two inventions can enlarge the speech communication areas to further quiet and noisy environments, and for young and aged people including speech or hearing handicapped people.

NAM is an extremely small voice, which is audible only for a speaker himself but inaudible for a nearby listener. NAM and a NAM microphone make it possible to communicate in the quiet environment, which is called a quiet telephone. NAM is also applied to speech recognition, which is called quiet speech recognition. NAM is also applied to speech handicapped aid using our voice conversion technique.

SIMO-ICA has the blind source separation (BSS) capability without distortion by the single-input multiple-output constraint. SIMO-ICA realizes hands-free communication in noisy environments. In order to implement realtime communication, we combined SIMO-ICA with the binary masking successfully.

We have been operating speech guidance systems in the real environments more than five years, based on large vocabulary speech recognition program (Julius). Takemaru-kun agent based speech guidance system has been used in Ikoma city north communication center. Takemaru-kun is positively accepted by Ikoma citizen, especially children. We have been also operating two types of speech guidance system in the noisy environment at the nearby railway station. They are a robot type (Kita-robo) and an agent type (Kita-chan). These two speech guidance systems are accepted by the railway company and wide range of users. We combine successfully BSS with Kita-robo in the realistic robot dialog research room to show the reality of the hands-free speech dialog system.

Keyword Speech media, Universal communication, Non-audible murmur(NAM), Blind source separation(BSS), Speech dialog system, Speech information guidance system, Robot dialog

1. まえがき

音声は、コミュニケーション手段として、人にとってもっとも自然なメディアである。ユニバーサルコミュニケーションを広げるには、コミュニケーションのギャップをなくする技術がさらに必要であり、どのような環境でも、誰もが、誰にでも自由にメディアが使える、人にやさしい技術の開発が重要となる。

奈良先端大情報科学研究科の音情報処理学講座では、音声と音を融合した分野の研究を進めている。この講演では、音声によるユニバーサルコミュニケーションの観点から研究室の最近のトピクスを取り上げて紹介する。

研究室で行ってきた音声メディアによる研究をユニバーサルコミュニケーションの観点からのまとめたものを図1に示す。静かな音声メディア「非可聴つぶやき(NAM)の発見[1,2,3]は、声を出さない音声コミュニケーションの可能性を示し、また、発話障害者の発話補助などへの応用の可能性[7,8]につながっている。

ブラインド音源分離における歪みなしでの分離の制約(SIMO-ICA)[9,10]は、高品質の音源分離の可能性を示し、かつ、バイナリーマスキングの併用による実時間処理[12,13,14]は、騒音下でのハンズフリー通話や両耳補聴器の可能性につながっている。

研究室では、大語彙連続音声認識プログラムJulius[15]を利用した音声情報案内システム「たけまるくん」を生駒市北コミュニティセンターに2002年に設置[16]して、5年以上に渡り運用している[18,20]。集録した幼児、子どもの音声データにより、幼児でも利用できる音声情報案内システムとなっている。

最近、ロボットとの音声対話を目指した雑音のある中でのハンズフリー音声認識を、ブラインド空間サブトラクションアレー(BSSA)[11]と音韻モデルと言語モデルによるデコーダーの結果を利用したデコーダーVAD[21,22]による実装ができ、良好な動作を実証した。

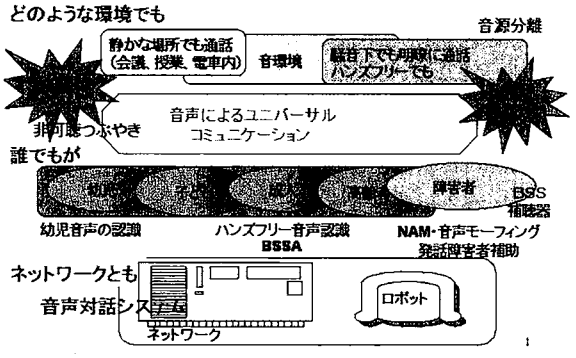

図1　奈良先端大における音声メディアによるユニバーサルコミュニケーションの研究

2. 非可聴つぶやき（NAM）

中島らによる静かな音声メディア「非可聴つぶやき(NAM)の発見[1,2,3]によるNAM発話とNAMマイクロホンの装着位置を図2に示しておく。NAMマイクロホンの開発[4]は、NAMによる電話（無音声電話）[5]やNAMによる認識（無音声認識）[1,2,3]を可能にした。声を出さない音声メディアの発見は、音声メディアの利用範囲を大きく広げるものであり、会議中や車内での静かな環境での音声コミュニケーションの可能性を提示している。

無音声電話の性能は、NAMから音声へのモーフィング[6]によって改善できた。このNAMから音声へのモーフィングの学習には、50発話程度の同じ発声内容のNAM発話と通常発話が利用されている。このモーフィングをGMM(Gaussian Mixture Model)に基づく統計的な枠組みで学習している[5]。NAM発話では、声帯が振動していないのでピッチ情報は抽出できないので、通常発話よりも、発声方法が近いささやき声への変換が容易であることも確かめられている。

NAM発話による音声認識、いわゆる無音声認識では、音韻モデルをNAM発話に適用することにより、ディクテーションなどの大語彙連続音声認識も可能であることが示されている[1,2]。音韻モデルを個人の発話に適応するには、50から数100 NAM発話による通常発話からの教師あり話者/発話適応MLLRが有効であり、明瞭度も高いことが確認されている。

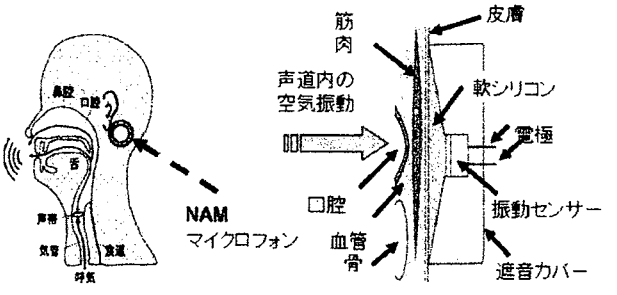

図2　非可聴つぶやき用NAMマイクロホン[4]

また、手術で声帯を失った発話障害者の発話補助の研究を行っている[7,8]。NAMマイクが口の中の小さな共振を捉える性能を利用して、微小振動子による口腔内共振をNAMマイクで捕らえ、ささやき声への変換（モーフィング）することにより、発話障害者の発話補助の可能性を見出している（図3）。

外部で聞こえない静かな音声メディア「非可聴つぶやき」の発見とNAMマイクロホンの開発は、図4に示すように様々なNAMの応用の可能性を示している。

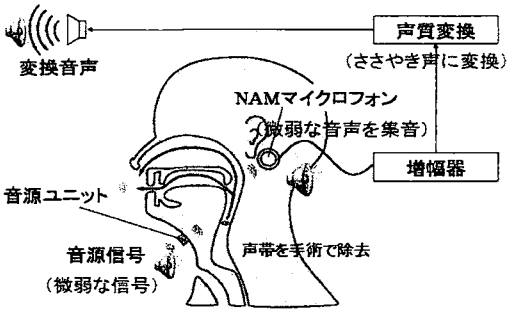

図3 微小振動子とNAMマイクロホンを利用した発話障害者補助[7]

図4 非可聴つぶやきとNAMマイクの応用の例

3. ブラインド音源分離

ブラインド音源分離は、複数の音源の独立性を仮定して、音源の分離を行う手法である。複数の音源の抽出フィルターを、音源の独立性だけを仮定して更新する教師なし学習で、独立成分分析（Independent Component Analysis: ICA）を用いている。通常の残響のある環境で音声の分離を行うには、残響が畳み込まれた音声の分離を行うので、スペクトル領域で周波数ごとに分離が行われる。スペクトル領域では、畳み込みが単純な掛け算になり、格段に分離が容易になる。しかし、周波数ごとに分離された成分を音源ごとに首尾一貫して分類する問題、いわゆるパーミュテーション問題(Permutation Problem)が生じ、この解決には、音源の到達方向(Direction of Arrival: DOA)や音源の調波構造などが利用されている。この手法は、ブラインド音源分離(Blind Source Separation: BSS)と呼ばれている。通常のBSSでは、音源の分離はできてSN比は向上するが、分離音には歪が含まれることが多い。

SIMO（Single-Input Multiple Output）拘束ICA(SIMO-based ICA)を導入することにより数学的には歪なしの分離が可能になった[9,10]。さらに、図5に示すように、両耳の分離音が歪なしで分離でき、音源の方向までも十分に保存されている。原理的に歪なしで分離できるので、音声認識への適用にも有効である。その他、両耳の音が分離抽出できるので両耳補聴器などへの応用も有望である。

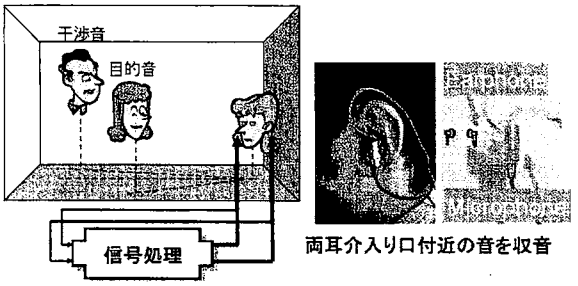

図5 SIMO-based ICAによるブラインド音源分離[9]

ブラインド音源分離は、統計的な分離アルゴリズムであるので、少なくとも3秒程度の入力が必要となる。一方、ハンズフリー通話のような音声によるコミュニケーションや音声認識では、実時間処理が不可欠である。一般に人や雑音の動きは、通常ゆっくりであるので、前の3秒で学習した分離フィルターを用いることが可能である。ただし、音声の始めは、他のアルゴリズム、あるいは、視覚情報による話者方向から推定した分離フィルターを用いることが必要となる。

我々は、実時間ブラインド音源分離には、図6に示すように、第1ステージはSIMO-base ICAブラインド音源分離、第2ステージにバイナリーマスキングの非線形処理を導入した[12,13,14]。企業と共同してDSPベースのブラインド音源分離装置を開発し、高い実時間音源分離性能を達成した。

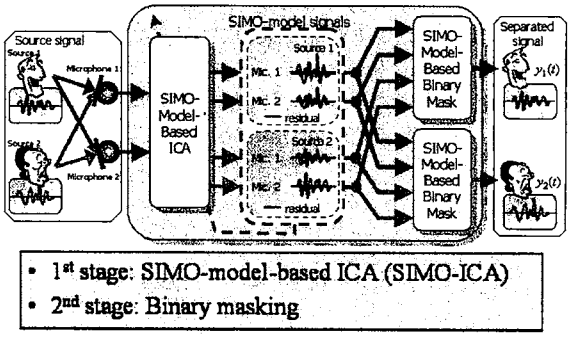

図6 SIMO-ICAとバイナリーマスキングによる実時間ブラインド音源分離アルゴリズム[12]

この手法は、基本的に500msec以下程度の低残響下で、かつ、人と人の声の分離のような点音源の場合には、極めて良好に動作する。拡散性の雑音と人の声の分離においても、一定の効果があることが実験的に確

かめられている。

ロボット対話などの応用では、背景の拡散性の雑音環境下でのハンズフリー音声認識が望まれる。このような状況では、近くの音声をBSSで抽出する性能より、背景の拡散性の雑音の方が、高いSN比で抽出されることが知られている。このような環境下で、遅延和アレーで強調された音声から、BSSを利用して抽出された雑音を減算する手法、BSSA(Blind Space Subtraction Array)が考案されている[11]。図7にBSSAのブロック図を示す。

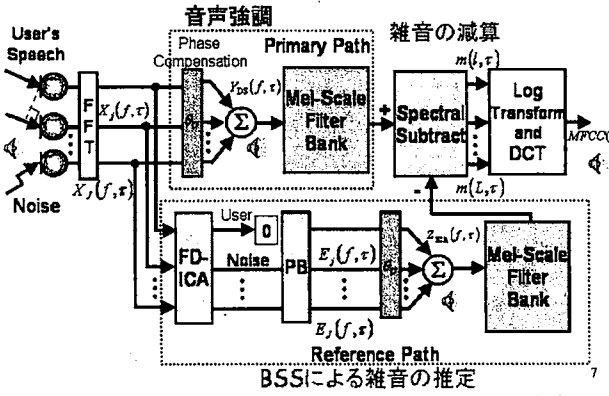

図7　BSSA(Blind Space Subtraction Array)による音声抽出のブロック図[11]

4. 音声情報案内システム

「たけまるくん」システム[16,18]は、平成14年11月から5年にわたり生駒市の北コミュニティセンターに常設され、情報案内サービスを行っている。マウスも利用できる。生駒市のキャラクターであるたけまるくんが、ユーザの質問に応答している。この音声情報案内システムは、図8に示すように、音韻モデル、言語モデル、質問応答データベースを備えている。大語彙連続音声認識プログラムJulius[15]を利用している。語彙数は4万語である。たけまるくんと生駒市北コミュニティセンターの外観を図9に示しておく。

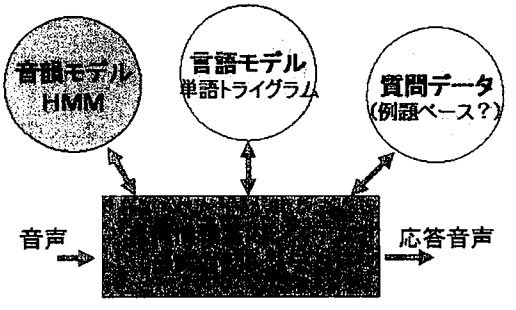

図8　音声情報案内システムの構成

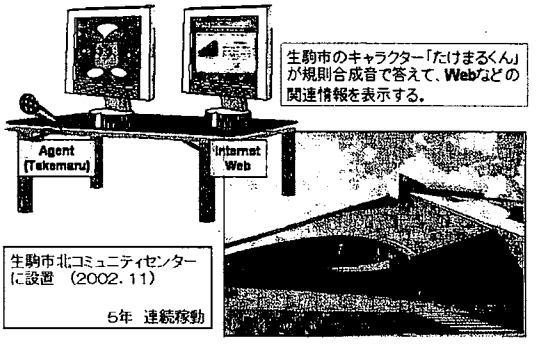

図9　たけまるくんと北コミュニティセンターの外観

当初の1年半は、300発話/日であったが、平成16年3月末に以下の改良を行い、600発話/日まで利用頻度があがっている。大人と子供のそれぞれの音韻モデルと言語モデルを利用した並列デコーディングにし、子どもの認識性能の向上と子ども向け応答文による改良を行った。さらに，笑い声，背景雑音などを発声長やGMMで検出できるようにした。音声応答は、一問一答型で、対話制御は行っていない。大量(約1万)の質問文例をもっており、各質問文例は、応答文と対応するWebアドレスをもっている。応答文の種類は，約500で、大人と子供別に用意されている。Juliusの認識結果から単語の確信度を計算して、語順を無視してもっとも近い質問文例を見つけ出し、対応する応答音声をTTSで合成し、かつ、Webの表示を行う。図10に月別の発話、笑い声、雑音の受付数を示しておく。1日あたりの入力の内訳を図11に示しておく。

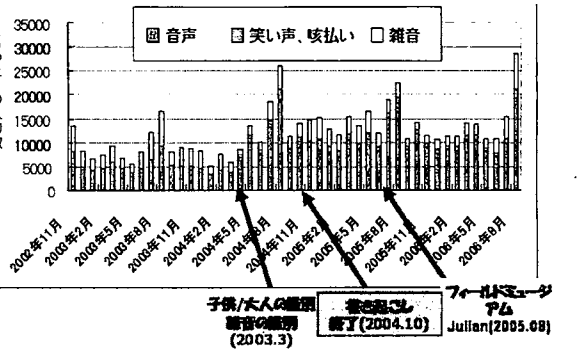

図10 たけまるくんの月別の発話,笑い声,雑音の受付数

現在の応答正解率は70%程度と推定される。誤った応答もニヤミスが多く、完全に間違えた場合でも、一問一答であるので、システムへの影響は少ない。「たけまるくん」は、生駒市のキャラクターであり、とくに子供に好かれており、大量の子供と幼児の音声データが収録できている。当初の子供単語認識率は、60%程

度であったが、現在では、85％程度までに向上している。また、幼児の認識は、まったくできなかったが、現在では60％程度にまで向上している。収録した音声データに基づいて、音韻モデル、言語モデル、質問応答データベースの学習アルゴリズムによる改良を行っている[17,18,19,20]。

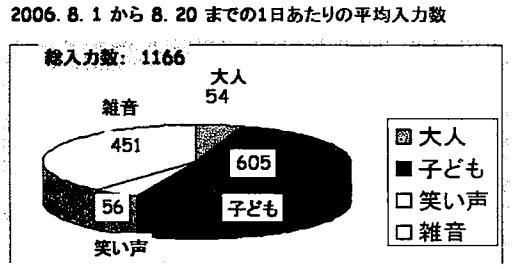

図 11 たけまるくんの１日あたりの入力の内訳

大学の近くの近鉄の学研北生駒駅に、２つのタイプの音声情報案内システムを 2006 年 3 月に設置した。エージェントタイプの「キタちゃん」とロボットタイプの「キタロボ」である。音声情報案内システムの「たけまるくん」からのポータビリティ、より騒音の高い環境での実証実験、ロボットタイプの効果などの検証を目指している。両システムとも大きな液晶ディスプレーを備えている。「きたちゃん」はタッチパネルも備えている。１日あたり 100 発話ほどの入力を受け、両システムとも当初から良好に動作し、たけまるくんとほぼ同様の性能を示している。ロボットタイプの「キタロボ」のほうが多くの音声入力を受け付けている。図 12 に１日あたりの入力の内訳を示しておく。

この音声データも集録され、半年間分の書きおこしも終了しており、「たけまるくん」からのポータビリティの観点から、音韻モデル、言語モデル、質問応答データベースの評価を行っている[18,20]。

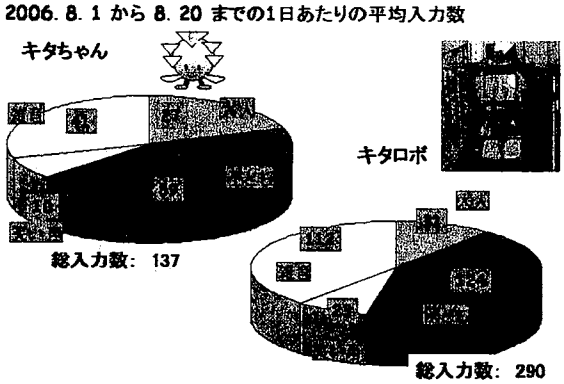

図 12 「キタちゃん」「キタロボ」の１日あたりの入力の内訳

5. ハンズフリーロボット対話

実環境でのハンズフリー音声認識を目指して、ハンズフリーロボット対話実証システムを構築している。

実環境でハンズフリーの開発や実験を行うことには様々な制約がある。実験室内で実環境とほぼ同じ音環境を多数のスピーカーで騒音レベルと騒音の方向を制御して再現し、かつ、残響時間も吸音パネルを用いて制御している。ハンズフリー音声認識システムを構築するには、少なくとも以下の技術が必要である。

（１）高 SN での実時間音声収録技術、

（２）騒音下での音声検出技術、

（３）雑音、残響、タスクに合った音韻モデル

（１）として、広がりをもつ背景雑音を除去できるブラインド空間サブトラクションアレー(BSSA)[11]の実時間処理を用いた。８チャンネルのマイクアレーを用いて、遅延和アレーで強調された音声から背景の広範囲の雑音を BSS で抽出してスペクトル減算を行っている[11]。（２）の騒音下での音声検出には、Julius 4 [25] デコーダーの中での単語検出により音声を切り出すデコーダーVAD[21,22]を用いた。このデコーダーVAD は、文頭のサイレンス（雑音）モデル、HMM 音韻モデル、さらに言語モデルの情報も利用しており、従来の雑音 GMM と音声 GMM を用いた手法よりもはるかに高い切り出し性能を示している[21,22]。この切り出しは、Julius4[23]に実装すみである。（３）の音響モデルには、たけまるくんの音韻モデルとキタロボ・きたちゃんで収録した音声にインパルスレスポンスを畳み込んだ音声で、MLLR-MAP を行ってタスク、残響、雑音にマッチした音韻モデルを構築している。図 13 にロボット対話実験室の様子を示す。

実験室で 60dBA の北生駒駅の騒音をながし、約 500msec の残響時間で、キタチャンタスクで呼び実験を行った。５名の発声者が約２ｍ離れた位置から発声し、切り出しの精度も含めた評価で、90%以上の単語認識精度を達成している。

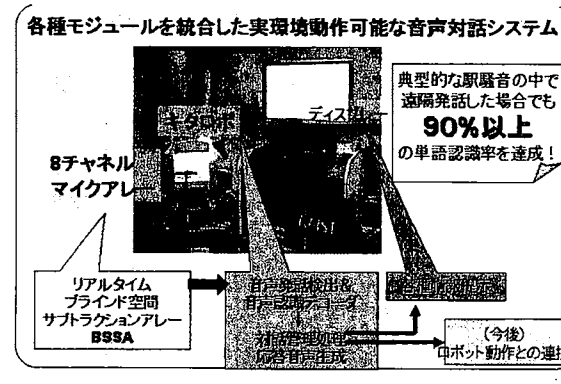

図 13 ロボット対話実験室

6. まとめ

研究室で行ってきた最近の音声認識、多チャンネル信号処理、音声合成の研究を、音声におけるユニバーサルコミュニケーションの観点から整理した。今後、これらの技術を社会に還元して、ユニバーサルコミュニケーションの発展に寄与できることを目指している。

最後に、研究室での研究を支えている猿渡准教授、戸田助教、川波助教、中島博士はじめ、博士課程の学生、修士課程の学生に感謝します。

この研究は、文部科学省のリィーディングプロジェクト e-Society などの援助を受けて行われた。e-Society プロジェクトのメンバーにも感謝します。

文　献

[1] 中島淑貴、柏岡秀紀、ニックキャンベル、鹿野清宏、" 非可聴つぶやき認識", 電子情報通信学会論文誌, Vol.J87-D-II, No.9, pp.1757-1764, September, 2004

[2] Yoshitaka Nakajima, Hideki Kashioka, Nick Cambell, Kiyohiro Shikano, "Non-Audible Murmur (NAM) Recognition", IEICE Trans. Information and Systems, Vol.E89-D, No.1, pp.1-8, 2006

[3] Y.Nakajima, H.Kashioka, K.Shikano, N.Campbell, "NON-AUDIBLE MURMUR RECOGNITION INPUT INTERFACE USING STETHOSCOPIC MICROPHONE ATTACHED TO THE SKIN", Proceedings of ICASSP2003, Vol.V, pp.708-711, April 2003.

[4] 中島淑貴,鹿野清宏,"非可聴つぶやきをインタフェースとするコミュニケーションのためのソフトシリコーン型 NAM マイクロホンの開発",電子情報通信学会論文誌 D, Vol. J89-D, No.8, pp. 1802-1810, August 2006

[5] Tomoki Toda, Alan W Black, Keiichi Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory",IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp.2222-2235, Nov. 2007

[6] Tomoki Toda, Kiyohiro Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models", Proceedings of Interspeech2005, pp.1957-1960, Sept. 2005

[7] 中村圭吾, 戸田智基, 猿渡洋, 鹿野清宏, "肉伝導人工音声の変換に基づく喉頭全摘出者のための音声コミュニケーション支援システム", 電子情報通信学会論文誌, Vol. J90-D, No. 3, pp. 780-787, Mar. 2007

[8] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," Proceedings of Interspeech, pp. 1395-1398, Sept. 2006

[9] Tomoya Takatani, Tsuyoki Nishikawa, Hiroshi Saruwatari, Kiyohiro Shikano, "High-Fidelity Blind Separation of Acoustic Signals Using SIMO-Model-Based Independent Component Analysis," IEICE Trans. Fundamentals, Vol.E87-A, No.8, pp. 2063-2072, August 2004

[10] Tomoya Takatani, Tsuyoki Nishikawa, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind Separation of Binaural Sound Mixtures Using SIMO-Model-Based Independent Component Analysis," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2004), Vol.IV, pp.113-116, May 2004

[11] Y.Takahashi, T.Takatani, H.Saruwatari, K.Shikano, "PERMUTATOIN-ROBUST STRUCTURE FOR ICA-BASED BLIND SOURCE EXTRACTION", Proceedings of ICASSP2007, April 2007

[12] Y.Mori, H.Saruwatari, T.Takatani, S.Ukai, K.Shikano, T.Hiekata, Y.Ikeda, H.Hashimoto, T.Morita, "Blind Separation of Acoustic Signals Combining SIMO-Model-Based Independent Component Analysis and Binary Masking," EURASIP Journal on Applied Signal Processing, vol. 2006, Article ID 34970, 2006

[13] Yoshimitsu Mori, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, Takashi Hiekata, Takashi Morita, "Blind Source Separation Combining SIMO-ICA and SIMO-Model-Based Binary Masking," Proceedings of ICASSP, pp. V-81-84, May 2006

[14] Yoshimitsu Mori, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, Takashi Hiekata, Takashi Morita, "HIGH-PRESENCE HEARING-AID SYSTEM USING DSP-BASED REAL-TIME BLIND SOURCE SEPARATION MODULE", Proceedings of ICASSP2007,April 2007

[15] オープンソースの高性能汎用大語彙連続音声認識 Julius : http://julius.sourceforge.jp/

[16] R.Nisimura, A.Lee, H.Saruwatari, K.Shikano, "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," Proceedings of ICASSP2004, Vol.I, pp.433-436, May 2004

[17] T.Cincarek, T.Toda, H.Saruwatari, K.Shikano, "Utterance-based Selective Training for the Automatic Creation of Task-Dependent Acoustic Models," IEICE Trans. Information and Systems, Vol.E89-D, No.3, pp.962-969, 2006

[18] T.Cincarek, R.Nisimura, A.Lee, K.Shikano, "INSIGHTS GAINED FROM DEVELOPMENT AND LONG-TERM OPERATION OF A REAL-ENVIRONMENT SPEECH-ORIENTED GUIDANCE SYSTEM",Proceedings of ICASSP2007, April 2007

[19] Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, ``Cost Reduction of Acoustic Modeling for Real-Environment Applications Using Unsupervised and Selective Training", IEICE Trans. Information and Systems, March 2008

[20] Tobias Cincarek, Hiromichi Kawanami, Ryuichi Nishimura, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano, ``Development, Long-Term Operation and Portability of a Real-Environment Speech-oriented Guidance System", IEICE Trans. Information and Systems, March 2008

[21] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano, A. Lee,"Voice Activity Detection Applied to Hands-Free Spoken Dialogue Robot based on Decoding using Acoustic and Language Model", 2007 International Conf. on Robot Communication and Coordination (ROBOCOMM2007), Oct. 2007

[22] 酒井啓行, ツィンツァレク トビアス, 川浪弘道, 猿渡洋, 鹿野清宏, 李 晃伸,"音響モデルと言語モデルに基づく音声区検出を用いたハンズフリー音声認識アルゴリズムの評価", 第 9 回音声言語シンポジウム、Dec. 2007

[23] 李晃伸、"大語彙連続音声認識エンジン", Julius ver.4", 第 9 回音声言語シンポジウム、Dec. 2007