様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 3 月 25 日現在

機関番号：13904
研究種目：若手研究（Ｂ）
研究期間：2009～2011
課題番号：21700192
研究課題名（和文）
プライバシー情報を隠蔽するための音声処理に関する研究
研究課題名（英文）
Research of speech signal processing for privacy protection
研究代表者
山本　一公（YAMAMOTO KAZUMASA）
豊橋技術科学大学・大学院工学研究科・助教
研究者番号：40324230

研究成果の概要（和文）：音声は多くのプライバシ情報を含むため、音声に含まれる一般に有用な情報を活用するためには、音声信号からプライバシ情報を取り除くことが条件となる。そのため、音声信号からプライバシ情報を取り除くための研究を行った。プライバシ情報を取り除くための基礎技術である、音声除去・音源分離、声質変換のための話者認識、遠隔発話音声認識のそれぞれにおいて、高精度に処理を行うための基盤技術を開発することができた。

研究成果の概要（英文）：Speech signal includes a large amount of privacy information. Removing of the information is necessary for using the general useful information included in the speech signal. We developed speech elimination / speech separation, speech recognition for voice conversion, distant speech recognition techniques as the fundamental techniques for removing the privacy information accurately.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,800,000	540,000	2,340,000
2010 年度	900,000	270,000	1,170,000
2011 年度	600,000	180,000	780,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学、知覚情報処理・知能ロボティクス
キーワード：音声情報処理、プライバシ保護、音声認識、話者認識、実環境、遠隔発話、音源分離

１．研究開始当初の背景

最近、インターネットの発達および安価なカメラの登場により、ストリートライブ映像配信等の臨場感通信を目的として、様々な場所にカメラが設置され、インターネットで配信されるようになってきた。これは、プライバシの問題を多分に含むと考えられるが、一般の人々はカメラ設置を社会安全のためであるとして許容しており、プライバシが侵害されているとはあまり感じていないようである。臨場感通信を目的とする場合、カメラには死角があり、暗くなると何も写らないという欠点があるため、そのような状況では、音情報を併用することが有用であると考えられる。

しかし、研究代表者の経験した事例（商業施設へのマイクロホン設置依頼）にもあったが、マイクは一般の人々がプライバシの面からその設置に対してカメラより敏感になっており、限定された空間であってもほとんど

設置されていない。

通常、カメラの映像は設置者自身の管理下に置かれるが、その制約を解いてセンサ情報を WWW のように共有しようという研究(「センシング Web プロジェクト」)がある。この研究では、様々なセンサ情報（例: カメラによる映像、温湿度センサによる温度・湿度等）を、プライバシ処理を施した上で一般に開放し、より広い範囲で利活用を図ることを目的としている。将来来るであろうユビキタスセンサーネットワーク社会において、音センサもその一翼を担うと考えられるが、カメラ設置よりもマイク設置の方がプライバシ侵害の度合いが大きいと考える人が多いため、プライバシ処理は画像よりも音情報に対して行うことが重要であると考えられる。

２．研究の目的

本研究では、プライバシを考慮して、収集した音情報からプライバシ情報を削除した上で、センサ情報として活用できる形に変換するシステムを開発することを目的とする。

マイクを設置して得られる音情報は、(1) 背景音（環境音)、(2) 音声（人の声）に大別できると考えられる。背景音には、自動車の走行音、人ごみの騒音、BGM などが考えられるが、これらはプライバシ上問題となることはないと考えられるため、プライバシ情報ではない情報として、そのまま利用できることが望ましい。このために、背景音と音声を分離する（音声除去）技術が必要となる。

一方、音声は、① 喋っているのは誰か（話者性）ということと、② 喋っている内容そのものがそれぞれプライバシ情報と成り得る。①に対しては、話者性を取り除いた音声信号を得るために声質変換技術が必要となる。プライバシ情報を隠蔽するために、ある一人の話者の声に全ての声をマッピングするのであれば、既存技術が存在するが、話者の区別ができなくなる。本研究においては複数人の会話の場合には複数人が会話していることを情報として保持する必要があるため、会話人数と話者の相違を識別して個別に声質変換を行う必要があり、声質変換に話者交代検出(話者認識を含む)を併用する必要がある。② に対しては、喋っている内容からプライバシ情報を除去するためには、音声からプライバシ情報と成り得る部分を特定し除去する（無音や``ピー音"に置換する）ことが必要となる。そのために、高精度な音声認識技術（話者はマイクから離れたところで喋るため、雑音に頑健な遠隔発話実環境音声認識技術）が必要となる。先の音声除去技術を開発できれば、それを転用することで雑音を除去することも可能になり、音声認識精度を向上させることができる。

３．研究の方法

(1) 音声除去と音源分離

① ベクトル量子化を基とする手法

図 1.に、本研究で提案したベクトル量子化(VQ)を基とした音源分離手法の基本的な構成を示す。本手法では、“背景音(図 1.では音楽)と音声の混合音”と“背景音のみ”のスペクトルをペアとして VQ により予めコードブックを作成しておく。背景音を含む入力音声が入力されると、コードブックからそれに最も良くマッチする混合音スペクトルを探し、そのコードに対応する背景音スペクトルを入力音声のスペクトルから差し引くことで、音声のみを抽出する。コードブックを作成するときに用いる背景音を別の背景音に変える(例えば、音楽を雑踏に変える)ことにより、別の環境に対応することができる。また、“背景音のみ”のスペクトルを“音声のみ”のスペクトルに入れ替えることで、音声を除去して背景音だけを抽出することが可能になる。

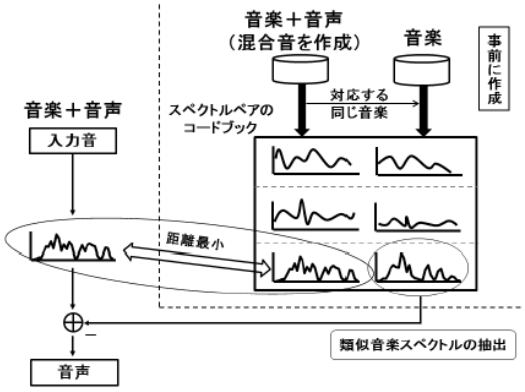

図 1. VQ を基とした音源分離

音声を抽出して提示または音声認識を行う場合には、上記手法をそのまま適用するよりも、コードブックを混合音+音声により構築し、「コードブックの音声スペクトル/コードブックの混合音スペクトル」のように Wiener フィルタ状のフィルタを構築して、入力音声をフィルタリングした方が高品質になる。

② 非負値行列因子分解(NMF)に基づく手法

NMF では、音声の振幅スペクトルが非負値であることから、スペクトルの時間-周波数系列を行列と見なして、これを音声成分と背景音成分に分離する。音声と背景音の混合信号に NMF を適用する場合は、音声、背景音のそれぞれのスペクトル集合が予め学習データから取得可能なことから、入力信号データに対して重みのみを求める手法が提案されている。これを用いることで、音声と背景音を分離することができるが、NMF は繰り返し演算により行列を分解するため、計算コスト

が高く、リアルタイム処理に向かない。そこで、本研究では、音声と背景音に対する基底ベクトルを事前に求めるだけでなく、混合音の代表ベクトルに対する重みも事前に求めておき、VQ による入力音声に最も近い代表ベクトルを探索し、それに対応する重みを用いてフィルタを構成することで、音声を抽出する。計算量は、コードベクトルの探索とフィルタリング処理のみとなるため、リアルタイム処理が可能である。

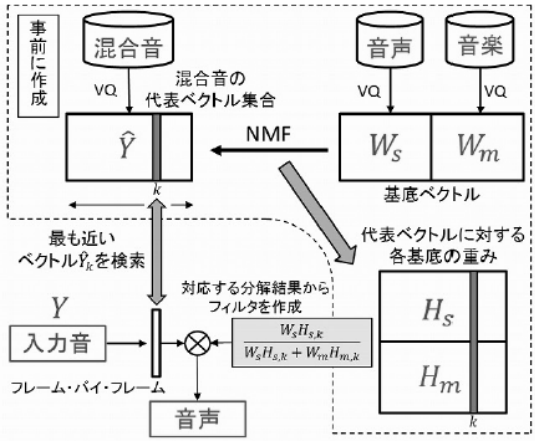

図 2. 高速 NMF

(2) 話者認識

従来の話者認識では、特徴パラメータとして MFCC(Mel-Frequency Cepstral Coefficients)を主として用いており、音声に含まれている位相情報は無視されている。位相は、音源波形の特徴によって大きく影響を受け、声道の形によっても影響される。これに基づき、位相情報を使った新しい特徴量の抽出法を考案し、それを用いて GMM を作成することで、テキスト独立の話者認識を行う方法を提案してきた。

サンプリングされた音声波形において N 個のサンプル点を切り出す。これを離散フーリエ変換することで以下の N/2 個の線スペクトルを得る。ここで、位相は時間と共に進行するため、切り出した波形が同じ周期波形であっても、切り出す時間位置によって位相が変化するという問題が生じる。また、周波数によって進行する角度が異なることも問題となる。この問題への対処として、ある基準とする角周波数の位相を一定にして、他の周波数における位相を相対的に求める手法を取った。具体的には、基準となる周波数 ω_b の位相を $\theta(\omega_b, t)$ とし、ω_b の位相が $\pi/4$ になるように、それぞれの周波数に応じて位相を次式により回転する：

$$\theta(\omega,t)+\frac{\omega}{\omega_b}\left(\frac{\pi}{4}-\theta(\omega_b,t)\right)$$

なお、本実験では基準周波数 $f_b = 1,000$ Hz とした。本手法により、位相は大まかに正規化されるが、基準周波数より低い位相は正しく正規化されない。しかし、基準周波数を低く取ると、その成分が弱く位相が信用できない場合が多々あるため、問題がある。そこで、位相正規化の精度を上げるために、分析窓の時間位置をピッチに同期させることを考える。正確なピッチ抽出は条件により難しい場合もあるので、ここでは分析窓内で最も信号振幅が大きい時刻を擬似ピッチマークと見なし、この擬似ピッチマークが分析窓の中心に来るように窓を数 ms 前後に動かすことによって、位相抽出の精度を上げる。

音声の位相は、背景雑音のレベルが高くなり SNR が悪くなると、背景雑音の位相により乱されるため、SNR が高い部分のみを用いることで、認識性能の向上を試みた。具体的には、スペクトルサブトラクション法と同じように音声の冒頭部分の雑音区間から雑音レベルを推定し、それを用いて SNR を算出、音声区間全体から SNR の低い部分を数十%程度削除して認識を行った。また、無声音区間では、声帯音源がランダム音源となるため、位相もランダムとなる。そのため、有声音区間を推定し、有声音区間だけを用いることで認識性能の向上を試みた。

(3) 遠隔発話音声認識

① マイクロホンアレイネットワークと話者位置・話者方向の同時推定

マイクロホンから離れた位置で行われる発話(遠隔発話)の認識を高精度に行うためには、音声を高品質に受音する必要がある。そのためには、話者の位置と発話方向を高精度に推定することが重要となってくる。本研究では、従来の多素子マイクロホンアレイを用いた方向推定で行われている解析的な手法ではなく、簡便なマイクロホンアレイを多数用いるマイクロホンアレイネットワークとニューラルネットワーク(ANN)による位置・方向の同時推定を提案した。概要を図 3. に示す。本研究では、ANN への入力として、マイクロホンアレイの各マイクロホンの位置(座標)、それぞれのマイクホロンアレイにおける音声信号の相関値とパワー、マイクロホン間の到達時間推定値(TDE)を用いている。

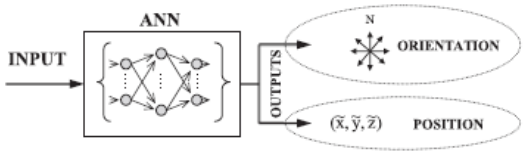

図 3. ANN による話者位置・方向の同時推定

② マイクロホンアレイネットワークによる遠隔発話音声認識

話者の位置・方向が推定できれば、その位置に対してマイクロホンアレイでビームを形成することにより、高精度な音声認識が可

能となる。しかし、マイクロホンアレイネットワークを設置する空間の特性に音声信号に歪が生じ、これが原因で音声認識性能が低下することが考えられる。そこで、従来から音声特徴量(ケプストラム)の伝達特性の正規化に用いられている CMN(Cepstral Mean Normalization; ケプストラム平均正規化)法を拡張して、GMM-based CMN を提案した。

従来の CMN では、発話内のケプストラムの全体の平均値を求め、それを伝達特性と見なして、各時刻のケプストラムから差し引く。発話が短い場合は、発話の内容に平均値が影響を受けるため、伝達特性のみを差し引くことが難しい。そこで、図 4.に示すように、GMM(Gaussian Mixture Mode; ガウス混合モデル)により、学習音声の音響空間(ケプストラム空間)を CMN したケプストラムの平均とともにクラスタリングしておき(クラス内の CMN する前のケプストラムの平均を μ_n、CMN したケプストラムの平均を $\tilde{\mu}_n$ とする)、入力音声のケプストラム C_i が最も近いクラス i の平均値 $\tilde{\mu}_i$ を用いて、

$$\tilde{C}_t = C_t - \frac{1}{T}\sum_{t=1}^{T}\left(C_t - \tilde{\mu}_{i(t)}\right)$$

とすることで正規化を行う。

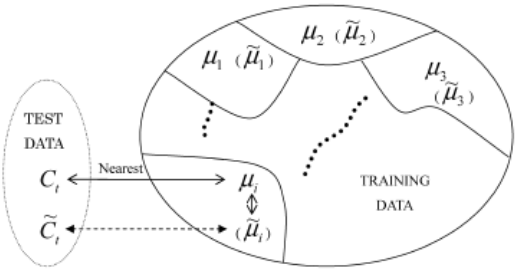

図 4. GMM-based CMN

4．研究成果

(1) 音声除去と音源分離

図 5.に VQ 手法により、音声を除去した背景音の聴取実験結果を示す。この実験では、音声除去を施すことで、元の音声と比べてどの程度単語が分かりにくくなったかを示している。

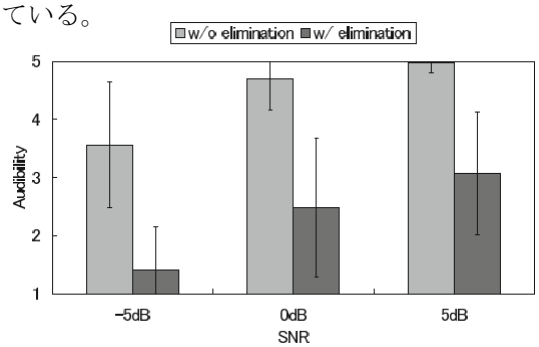

図 5. 音声除去実験結果

縦軸は可聴度で、横軸は音声と背景音の比率(SNR)である。可聴度は 5 段階で、5 が最も聴き取りやすく、1 が最も聴き取り難い。音声には新聞読み上げ音声データベースの音声を、背景音にはレストラン雑音を用いた。図から分かるように、音声除去を行うことで、単語の判別がかなり難しくなっていることが分かる。

次に、音源分離を適用した音声認識結果を示す。東北大・松下単語データベース(212 単語)を用いた結果で、音響モデルは背景音を含まないデータで学習したものである。提案法により、音声認識性能が改善していることが分かる。

表 1. 音源分離による音声認識結果
両手法の併用 (a:VQ+NMF, b:VQ+NMF 法の高速化)

入力・手法	SNR			
	−5dB	0dB	10dB	20dB
処理なし	2.2	7.8	53.4	86.1
VQ 手法	8.0	20.0	74.1	90.9
NMF 法	21.1	43.4	83.2	93.2
NMF 法(高速化手法)	5.2	17.6	71.4	90.4
両手法の併用 a	21.1	43.4	83.3	93.6
両手法の併用 b	8.0	21.9	74.7	91.8
クリーン音声	98.8			

(2) 話者認識

話者認識実験は、NTT データベースを用いて行った。従来の、MFCC を特徴量とする GMM ベースの話者認識手法では 97.1%の話者識別精度であったものが、MFCC を擬似ピッチ同期分析を導入した位相特徴量と組み合わせることで 98.7%に、さらに有声音区間のみを用いることで 98.73%まで識別精度が向上した。位相情報は、発話速度が速い、またはゆっくりな音声の話者識別に対して特に有効であった。

(3) 遠隔発話音声認識

遠隔発話音声認識の実験は、図 6.に示す 5m×6.4m×2.65m の部屋で行った。A～H の 8 カ所に T 字型マイクロホンアレイが設置されている(A～D は壁、E～H は天井に設置)。T 字型マイクロホンアレイは 4 素子で、マイクロホンの組み合わせを基準マイクとそれ以外で 3 つ持つため、ANN への入力は 136 次元である。音声データは、家電制御のためのコマンドや数字である。

話者位置推定については、平均誤り距離が 3 次元座標で 34.3cm、高さを無視した 2 次元座標で 29.4cm であった。また、8 方向の方向推定精度は、92.8%であった。

話者位置推定結果を利用した音声認識では、基準マイクのみで認識を行う場合の認識精度が 69%であったのに対して、スペクトルサブトラクションによる雑音除去と、推定位置情報を用いたビームフォーミングを組み合わせることによって、86.13%まで認識精度

が向上した。また、GMM-based CMN を併用することで、更に 87.90%まで認識精度を向上させることができた。

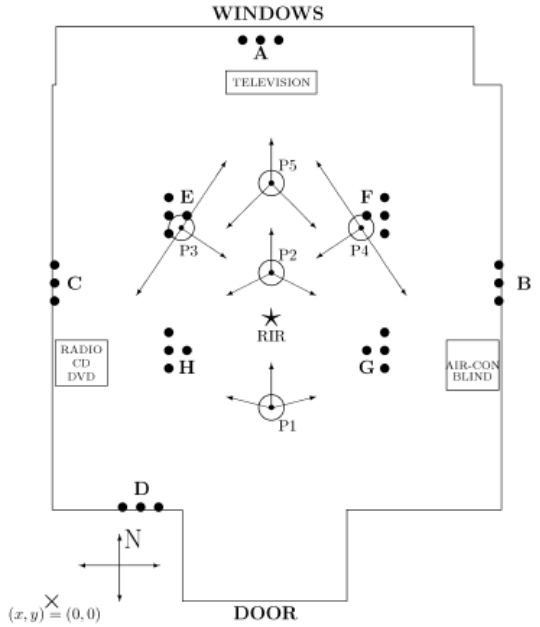

図 6. 実験環境

５．主な発表論文等

〔雑誌論文〕（計 7 件）

① Takahiro Fukumori, Takanobu Nishiura, Masato Nakayama, Yuki Denda, Norihide Kitaoka, Takeshi Yamada, Kazumasa Yamamoto, Satoru Tsuge, Masakiyo Fujimoto, Tetsuya Takiguchi, Chiyomi Miyajima, Satoshi Tamura, Tetsuji Ogawa, Shigeki Matsuda, Shingo Kuroiwa, Kazuya Takeda, Satoshi Nakamura, "CENSREC-4: An evaluation framework for distant-talking speech recognition in reverberant environments," Acoustical Science and Technology, Technical Report, Vol.32, No.5, pp.201-210, Sep. 2011.

② Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto, "Auditory perception versus automatic estimation of location and orientation of an acoustic source in a real environment," Acoustical Science and Technology, Vol.31, No.5, pp.309-319, Sep. 2010.

③ Longbiao Wang, Kazue Minami, Kazumasa Yamamoto, Seiichi Nakagawa, "Speaker recognition by combining MFCC and phase information in noisy conditions," IEICE Transactions on Information and Systems, Vol.E93-D, No.9, pp.2397-2406, Sep. 2010.

④ Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto, "Distant speech recognition using a microphone array network," IEICE Transactions on Information and Systems, Vol.E93-D, No.9, pp.2451-2462, Sep. 2010.

⑤ Kazumasa Yamamoto, Masatoshi Tsuchiya, Seiichi Nakagawa, "Privacy protection for speech signals," Procedia - Social and Behavioral Sciences, Vol.2, No.1, pp.153-160, 2010.

⑥ Kazumasa Yamamoto, Seiichi Nakagawa, "Privacy protection for speech information," Journal of Information Assurance and Security (JIAS), Vol.5, No.1, pp.284-292, 2010.

⑦ Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone array network," Journal of Acoustical Society of America, Vol.126, No.6, pp.3084-3094, Dec. 2009.

〔学会発表〕（計 19 件）

① 仲野翔一, 山本一公, 中川聖一, "音楽重畳音声の音声認識のための NMF による音楽除去の高速化および VQ 手法の改善", 日本音響学会 2012 年春季研究発表会講演論文集, 1-P-18, 神奈川大学, Mar.13, 2012.

② 嶋田晃太, 山本一公, 中川聖一, "残響に頑健な遠隔発話の話者認識の検討", 日本音響学会 2012 年春季研究発表会講演論文集, 1-P-29, 神奈川大学, Mar.13, 2012.

③ Kohta Shimada, Kazumasa Yamamoto, Seiichi Nakagawa, "Speaker identification using pseudo pitch synchronized phase information in voiced sound," Proc. 2011 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011), CD-ROM, Xi'an, China, Oct. 20, 2011.

④ 山本一公, 中川聖一, "長時間位相特徴と振幅スペクトル特徴の併用による音声認識の検討," 日本音響学会 2011 年秋季研究発表会講演論文集, 2-Q-13, pp.95-98, 島根大学, Sep. 21, 2011.

⑤ Shoichi Nakano, Kazumasa Yamamoto, Seiichi Nakagawa, "Speech recognition in mixed sound of speech and music based on vector quantization and non-negative matrix factorization," Proc. INTERSPEECH 2011, pp.1781-1784,

Florence, Italy, Aug. 29, 2011.
⑥ 仲野翔一, 山本一公, 中川聖一, “NMF と VQ 手法による音楽重畳音声の音声認識,” 電子情報通信学会技術研究報告, SP2011-34, pp.23-28, 名古屋大学, Jun. 23, 2011.
⑦ 嶋田晃太, 山本一公, 中川聖一, “有声音部の位相情報を用いた話者認識の改善,” 日本音響学会 2011 年春季研究発表会講演論文集, 2-5-4, pp.51-54, 早稲田大学, Mar. 10, 2011.
⑧ 仲野翔一, 山本一公, 中川聖一, “NMF と VQ 手法による音楽重畳音声の音楽除去と音声認識,” 日本音響学会 2011 年春季研究発表会講演論文集, 2-P-14, pp.159-162, 早稲田大学, Mar. 10, 2011.
⑨ Yasuhisa Fujii, Kazumasa Yamamoto, Seiichi Nakagawa, “Large Vocabulary Speech Recognition System: SPOJUS++,” Proc. 11th WSEAS International Conference on MULTIMEDIA SYSTEMS & SIGNAL PROCESSING (MUSP ’11), pp.110-118, Venice, Italy, Mar. 8, 2011.
⑩ Kazumasa Yamamoto, Eiichi Sueyoshi, Seiichi Nakagawa, “Speech recognition using long-term phase information,” Proc. INTERSPEECH 2010, pp.1189-1192, Makuhari, Japan, Sep. 28, 2010.
⑪ 山本一公, 末吉英一, 中川聖一, “長時間分析に基づく位相情報を用いた音声認識の検討,” 電子情報通信学会技術研究報告, SP2010-40, pp.31-36, 仙台 秋保温泉 緑水亭, Jul.23, 2010.
⑫ Longbiao Wang, Kazue Minami, Kazumasa Yamamoto, Seiichi Nakagawa, ″Speaker identification by combining MFCC and phase information in noisy environments,″ Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), pp.4502-4505, Dallas, USA, Mar. 16, 2010.
⑬ 嶋田晃夫, 山本一公, 中川聖一, “話者認識のための位相特徴抽出法の改善,” 日本音響学会 2010 年春季研究発表会講演論文集, 2-Q-19, pp.285-286, 電気通信大学, Mar. 9, 2010.
⑭ 末吉英一, 山本一公, 中川聖一, “長時間位相特徴パラメータによる音声認識の検討,” 日本音響学会 2010 年春季研究発表会講演論文集, 1-6-3, pp.9-10, 電気通信大学, Mar. 8, 2010.
⑮ 藤井康寿, 山本一公, 中川聖一, “大語彙連続音声認識システムの改善：SPOJUS++,” 第 4 回音声ドキュメント処理ワークショップ講演論文集, SDPWS2010-01, 豊橋技術科学大学, Feb. 26, 2010.
⑯ Kazumasa Yamamoto, Masatoshi Tsuchiya, Seiichi Nakagawa, “Privacy protection for speech signal,” Proc. International Conference on Security Camera Network, Privacy Protection and Community Safety 2009 (SPC2009), Kiryu, Japan, Oct. 29, 2009.
⑰ Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto, “Distant speech recognition using spatial information estimated by a microphone array network,”, 日本音響学会 2009 年秋季研究発表会講演論文集, 1-R-22, pp.191-194, 日本大学郡山キャンパス, Sep. 15, 2009.
⑱ Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto, ″Estimating the position and orientation of an acoustic source with a microphone array network,″ Proc. INTERSPEECH2009, pp.1127-1130, Brighton, UK, Sep. 8, 2009.
⑲ Kazumasa Yamamoto, Seiichi Nakagawa, “Privacy protection for speech information,” Proc. The Fifth International Conference on Information Assurance and Security (IAS2009), pp.717-720, Xi’an, China, Aug. 19, 2009.

〔図書〕(計 1 件)

① Kazumasa Yamamoto, Seiichi Nakagawa, “Evaluation of Privacy Protection Techniques for Speech Signals” in “Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications (Communications in Computer and Information Science)” E. Hüllermeier, R. Kruse, F. Hoffmann (Eds.), Springer, 2010, ISBN:978-3-642-14057-0.

６．研究組織

(1)研究代表者

山本　一公 (YAMAMOTO KAZUMASA)
豊橋技術科学大学・大学院工学研究科・助教
研究者番号：40324230

(2)研究分担者

なし

(3)連携研究者

なし