様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１１　日現在

機関番号：　３３９１０
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００２３３

研究課題名（和文）人間の聴覚特性を導入した深層ニューラルネットワークによる高精度な実環境下音声認識

研究課題名（英文）Accurate speech recognition system with deep neural network introducing human auditory characteristic in real environments

研究代表者

山本　一公（YAMAMOTO, Kazumasa）

中部大学・工学部・准教授

研究者番号：４０３２４２３０

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：現在、音声認識技術に深層学習が導入され、徐々に実用的に使われるようになってきているが、雑音環境下等での音声認識性能は未だ十分ではない。本研究の目的は、DNN (Deep Neural Network) 音響モデルに人間の聴覚特性を融合させることで、音声認識精度改善を得ることである。
本研究では、人間の聴覚特性を考慮した特徴抽出フィルタバンクを深層学習により自動的に学習する手法を提案した。この手法により、不特定話者音声認識に対する音声認識精度の改善を得た。また、提案手法により適応化データ量が少ない条件下における話者適応化においても認識精度の改善が得られ、効果的であるという結果が得られた。

研究成果の概要（英文）：Currently, deep learning has been introduced into speech recognition technology and the speech recognition technology is gradually being used practically, but speech recognition performance is still not sufficient in noisy environments or for distant-talking. The purpose of this research is to improve speech recognition accuracy by combining DNN (Deep Neural Network) acoustic model with human auditory characteristics.
In this research, we proposed a method to automatically learn feature extraction filterbanks at the bottom of DNN acoustic model by using deep learning considering human auditory characteristics. By using this method, improvement of speech recognition accuracy was obtained for speaker-independent speech recognition. In addition, the proposed method improved speaker-adapted speech recognition accuracy even under the condition that the amount of adaptation data is small. The results showed the effectiveness of the proposed method.

研究分野：　音声情報処理

キーワード：　音声認識　深層学習　Deep Neural Network　聴覚特性　音響特徴量　フィルタバンク

１．研究開始当初の背景

音声認識技術は、特別な訓練を必要としないこと、マイク以外の特別な装置を必要としないこと等の利点により、コンピュータシステムへの自然な情報入力手段として古くから期待されており、その研究も長きに渡り行われて来た。ここ最近になって、Google「Google Now」、Apple「Siri」、NTT ドコモ「しゃべってコンシェル」等の実用的なサービスが展開されるようになり、ようやく一般の人々が音声認識技術を利用できるようになってきた。しかし、音声認識精度という観点から見ると、いつでもどこでも誰でも確実に認識してくれる訳ではなく（単語認識精度は 60%程度と言われている）、まだまだ音声認識性能の改善が必要な状況である。

最近の音声認識精度の改善は、音声認識技術に深層ニューラルネットワーク（Deep Neural Network; DNN）が導入されたことが大きい。これは、Hinton らの研究成果により DNN が精度良く学習できるようになったこと[①]、GPGPU（汎用画像処理ユニット）を用いた高速な行列演算処理が比較的安価に利用可能になったこと、コンピュータ技術の進歩に伴って大規模な学習用データベースが利用可能になったことが重なることで起こったことであり、近年の音声認識技術開発の中でも非常に大きなブレークスルーとなった。

ニューラルネットワーク（NN）を音声認識に用いる利点は、NN が持つ強力な汎化能力（学習データに出現しないテストデータに対する認識性能の高さ）であるが、静かな環境で丁寧に読み上げられた音声（クリーン音声）のみを用いて学習した NN を用いて、雑音や残響のある実環境で発声された音声を認識するというような極端なタスクにおいても高精度に認識が可能という訳ではない。そのため、現在、世界中で NN を用いた実環境下での高精度な音声認識のための研究が行われている。代表的なものとしては、denoising autoencoder と呼ばれる雑音を含む音声特徴量をクリーン音声の特徴量に変換する方法[②]や、NN の中間層のノード数を少なくすることでそのノードの出力を次元圧縮された特徴量として用いるボトルネック特徴[③]がある。また、人間の神経ネットワークが局所的に結合されていることに注目し、これを模擬した畳み込み NN（CNN）[④]も広く研究されている。

２．研究の目的

本研究では、DNN を用いる音声認識技術において、NN が模擬していると考えられる（暗黙的に学習される）人間の神経ネットワークに、人間の聴覚神経に対応するネットワークを明示的な事前知識に基づいて組み込むことによって、より効果的に音声認識精度の改善を図ることを目的とした。人間の聴覚神経に対するネットワークとして具体的に、従来特徴抽出に用いられている対数的周波数分解能や対数的振幅圧縮特性、等ラウドネス特性、臨界帯域特性だけでなく、マスキング特性（ある時刻ある周波数の成分がマスカーとなり、他時刻他周波数の成分を抑圧する特性）、周波数帯域選択的聴取による雑音低減能力、両耳聴取による雑音低減能力、一次聴覚野における時間-周波数特性を明示的に導入することで認識精度を改善することを試みる。もちろん、NN の万能性から、事前知識を与えずとも、音声認識に必要と考えられる特徴量が NN 内部において暗黙的・自動的に学習される可能性は十分にある。しかし、通常、そのような特徴が自動的に抽出されることを期待するためには、十分に深いネットワーク、十分に多い隠れノード数、それらを学習するために十分に多い学習データが必要となるため、現実的にこれは難しいと考えられる。そのために、我々が持つ事前知識を NN 上に明示的に再現することによって、認識性能の改善を効率的に得る技術を開発することを研究目的とした。

具体的には、「NN による人間の聴覚特性の効率的な表現方法」、「聴覚特性を表現するために適した NN の構造検討」について研究を行った。

３．研究の方法

(1) 研究の当初の構想としては、DNN の入力層に近い特徴抽出層の係数を、聴覚フィルタの係数で初期化することで、精度の向上を図る予定であった。しかしながら、そのように DNN の初期値を設定しても、ランダムに初期値を設定しても、DNN の係数は上書き学習されてしまい違いが生じないことが分かった。その後、学習可能なパラメトリックフィルタを DNN の特徴抽出層に用い、このフィルタパラメータを自動的に学習することで、より音声認識に適した特徴抽出を自動的に行うことができる DNN に対しての研究を行うこととした。

これまで行われてきた特徴抽出フィルタバンクを自動的に学習する手法では、フィルタ形状そのものを明示的に与えず、結果として得られるフィルタ形状を分析することで、音声認識に適した特徴抽出が行われているかどうかを事後的に検証する方法が採られてきた。そこで本研究では、文献[⑤]で提案されたフィルタバンクの設計方法を DNN の枠組みに組み込み、音響モデルとともにフィルタバンクを統一的に学習する手法を提案する。

フィルタバンクを組み込んだ DNN の外観を図１に示す。従来の音声認識用音響特徴抽出では、人間の事前知識に基づくフィルタバンク（メル周波数軸上で等間隔に配置された三角フィルタバンク）を用いて特徴抽出が行われてきたが、本研究ではこれをパラメトリックフィルタに置き換え、DNN の他のパラメータと同じ枠組みの中で自動的に学習する。入

力となるパワースペクトルは、複数フレームを同時に入力するが、フィルタはフレーム間で共通とし、パラメータを共有した上で特徴抽出処理および学習によるパラメータ更新処理を行う。DNN の出力層は音素単位の隠れマルコフモデル（Hidden Markov Model; HMM）の状態に対応しており、softmax 演算を行うことで、状態の事後出力確率を計算することができる。

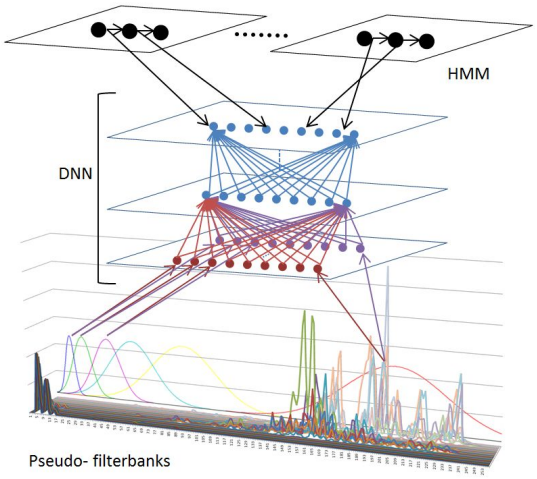


図 1．フィルタバンクを統合した DNN の概観

フィルタバンクに用いるフィルタの形状は、ガウス関数を用いたもの:

$$\theta_n(f) = \varphi_n \exp\left\{-\beta_n(p(\gamma_n) - p(f))^2\right\}$$

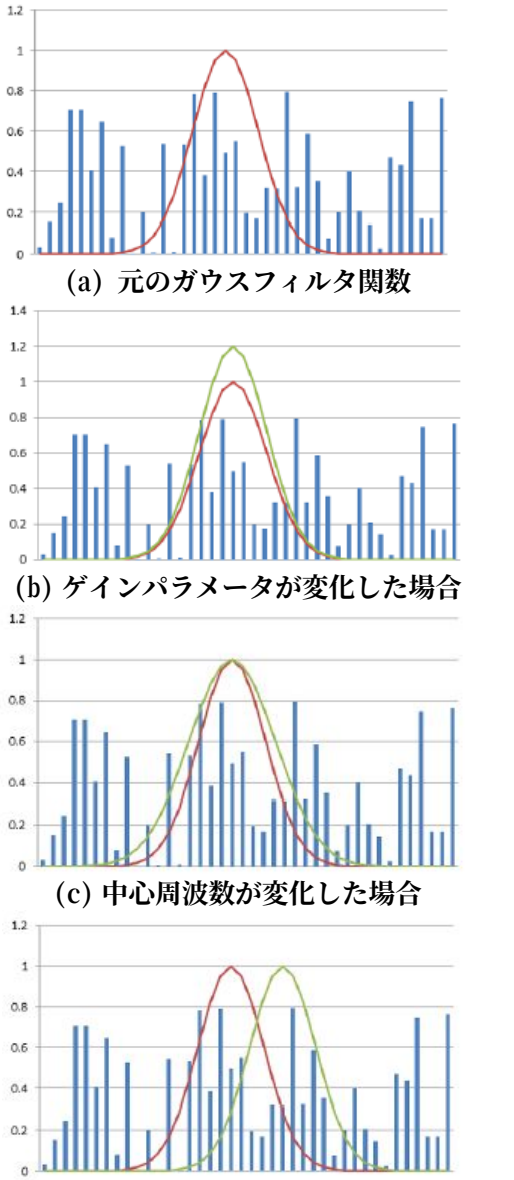
（フィルタ 1 つ当たりのパラメータ数はゲイン φ_n、中心周波数 γ_n、バンド幅 β_n の 3 つ）と、ガンマトーンフィルタを用いたもの

$$\theta_n(f) = |H_n(f)|^2 \sim k_n^2 \left\{\left[1 + \frac{(f - f_0(n))^2}{b_n^2}\right]^{-a} + \left[1 + \frac{(f + f_0(n))^2}{b_n^2}\right]^{-a}\right\}$$

（フィルタ 1 つ当たりのパラメータ数はゲイン k_n、中心周波数 $f_0(n)$、時間減衰係数 b_n の 3 つ）の 2 種類について検討を行なった。3 つのパラメータの変化に応じてフィルタ形状が変化する様子を図 2 に示す（ガウス関数フィルタの場合）。図 2(a)はパラメータが更新される前のガウスフィルタ関数であり、3 種類のパラメータが更新されることによって、図 2(b)(c)(d)のようにフィルタ形状が変化する。これによって、より音声認識に適したフィルタ形状へと自動的に学習される。

それぞれのフィルタにおいて、バックプロパゲーションアルゴリズムに基づくパラメータ更新式を導出し、DNN のその他のパラメータと共に統一的に自動学習を行う。

(2) 提案モデルは、フィルタバンク層が少ないパラメータ数で構成されるといった利点があり、環境適応化や話者適応化しやすいモデルであると考えられる。表 1 に、実験で使用したモデルの更新対象パラメータ数を示す。表中、“提案法（ガンマトーンフィルタ）”は提案モデルにおいてガンマトーンフィルタを用いた場合、“fDLR”[⑥]はフレーム間で重みを共有する線形層を DNN の最下層に挿入する手法（DNN への入力は三角フィルタバンク出力）、“ExpFDNN”[⑦]は、入力となるパワースペクトルの周波数ビン毎の指数関数重みを制約無しで推定する手法である。実験ではフィルタの総数を 40 としたため、提案法でフィルタバンク層を学習する場合のパラメータ数は 120(3 パラメータ×40 フィルタ)である。一方、線形層を学習する fDLR の更新対象パラメータ数は 1600（40 入力×40 出力の行列）、ExpFDNN ではパワースペクトルの周波数ビン数が 256 であるため更新対



(a) 元のガウスフィルタ関数

(b) ゲインパラメータが変化した場合

(c) 中心周波数が変化した場合

(d) バンド幅が変化した場合

図 2．フィルタにおけるパラメータの役割

象パラメータ数は 10,240（256 周波数ビン×40 フィルタ）であり、提案モデルが少ないパラメータで構成されていることが分かる。このことから、提案手法は少ないデータで行う環境適応化および話者適応化に適した手法であると考えられる。

表 1．更新対象パラメータ数の比較

フィルタ	パラメータ数
提案法(ガンマトーンフィルタ)	120（3×40）
fDLR	1600（40×40）
ExpFDNN	10,240（256×40）

4．研究成果

(1) フィルタの自動学習により音声認識性能が向上することを示すために、不特定話者音声認識実験を行なった。

データベースとして、ASJ データベースおよび JNAS データベースを用いた。データベースは男性話者と女性話者で構成されており、これら 2 つのデータベース(性別データ)を個別に用いて 2 つのモデルを別々に学習する場合と、両データベースを同時に用いてひとつのモデルを学習する場合の、計 3 モデルを学習する。また、これ以降男性話者は SM (speaker male)、女性話者は SF (speaker female) と表記する。学習データ量は男女話者それぞれ 33 時間と 44 時間、話者数はそれぞれ 133 話者と 164 話者である。評価は、男性話者 100 発話(IPA100 文、23 話者)、女性話者 100 発話(23 話者) を用いて行う。いずれのテストセットも未知語率は 0.4 %、パープレキシティは 125.7 であった。

音響モデルは triphone DNN-HMM hybrid system とし、学習ラベルの強制アライメントのための GMM-HMM を最尤推定を用いて学習した。言語モデルは、毎日新聞データベースから学習した trigram で、語彙サイズは 2 万語である。GMM-HMM および言語モデルの学習、WFST デコーダは Kaldi 音声認識ツールキットを用いて学習・実装を行った。使用した DNN のアーキテクチャと音響特徴量は以下のとおりである。

ベースライン: 音響特徴量として、HMM Toolkit で求めた 40 次元メルスケールフィルタバンク出力(固定三角フィルタバンクを使用) を用いた。DNN への入力として、当該フレームと前後 5 フレームを連結した 11 フレーム、計 440 次元音響特徴量を使用した。ネットワークは 5 層、ユニット数は 2048 ユニットとし、出力層は triphone 音響モデルの状態数である 3,234 ユニットとした。

GFDNN: ガウス関数フィルタバンク層を持つ GFDNN (Gaussian filterbank incorporated DNN)を学習した。音響特徴量として、256 ビンパワースペクトルを使用した。GFDNN への入力として、当該フレームと前後 5 フレームを連結した 11 フレーム、計 2,816 次元音響特徴量を使用する。ネットワークは、フィルタバンク層を含む 6 層、第 1 隠れ層のユニット数は 440（ベースラインの入力層に合わせた)、それ以降は 2,048 ユニットとし、出力層は triphone 音響モデルの状態数である 3,234 ユニットとした。フィルタバンク層のフィルタ総数は、ベースラインのフィルタバンクと揃え 40 とした。フィルタパラメータの初期値は、ゲインを 1.0、中心周波数はメルスケール上で等間隔、バンド幅は 2σ区間が配置したメルスケールバンド幅になるように設定した。

GtFDNN: ガンマトーンフィルタバンク層を持つ GtFDNN (Gammatone filterbank incorporated DNN) を学習した。ハイパーパラメータは GFDNN と揃え、モデルの学習を行った。

ExpFDNN: 提案法との比較のために、使用した。パラメータ数の比較で示した通り、入力は 256 ビンパワースペクトル、フィルタ数は 40 とした。その他の条件は GFDNN に準ずる。

表 2．単語誤り率(WER)による性能比較

手法	Word Error Rate [%]			
	SM	SF	平均	SM+SF
ベースライン	4.9	5.0	5.0	5.0
GFDNN	4.1	4.7	4.4	4.1
GtFDNN	4.7	4.1	4.4	4.0
ExpFDNN	5.1	5.1	5.1	4.1

表 2 に不特定話者音声認識実験結果（単語誤り率）を示す。ベースライン（三角フィルタバンク）では、単語誤り率は男性で 4.9%、女性で 5.0%、平均で 5.0%であった。ベースラインはフィルタを学習しておらず、フィルタパラメータは固定である。フィルタを学習する提案手法である GFDNN と GtFDNN では、男女の平均単語誤り率は 4.4%となっており、ベースラインから大きな改善が見られた。学習により、フィルタ形状を学習することが重要であることが、この結果から分かる。男性・女性を別々にモデル化せず、学習データをまとめて性別に依存しないモデルを作成した場合（SM+SF）では、更に認識精度の改善が得られている。従来法との比較では、性別に依存したモデル（SM および SF）では、ExpFDNN の単語誤り率は 5.1%とベースライン手法とほぼ同一の結果となり、提案手法の効果が見られた。しかしながら、SM+SF では ExpFDNN の性能が向上し、提案手法と同程度の認識性能になっている。これは、ExpFDNN のフィルタ部分のパラメータ数が多いため、

学習データ数が増える SM+SF 条件において、モデルパラメータが十分に学習できた結果によるものと考えられる。このことから、適応文数が少ない話者適応化手法においては、提案手法の方が効果が高くなると考えられる。

(2) 提案モデルの話者適応化への効果を示すために、話者適応化を用いた音声認識実験を行なった。データベースとして、CSJ (Corpus of Spontaneous Japanese)の学会講演男性データを使用した。評価セットとして、CSJ 付属のテストセット 2 を使用した。話者は 5 名で構成されており、20 発話を適応セット、40 発話をテストセットに割り当てた。音響モデルおよび使用した DNN のアーキテクチャ、音響特徴量は(1)の実験と基本的には同じであるが、既存の話者適応化の手法として、fDLR を用いた適応化も行い、提案手法である GFDNN および GtFDNN との比較を行う。

fDLR:提案法との話者適応化における比較手法として使用した。入力は 40 次元の固定三角フィルタバンク出力であり、DNN の最下層に挿入される行列は 40×40 で、初期値は単位行列である。

表 3. 話者適応化による認識性能比較
（単語誤り率(WER) [%]）

適応文数	GFDNN	GtFDNN	fDLR	ExpFDNN
0	12.5	12.1	12.4	12.3
1	12.3	12.2	36.0	12.4
2	12.2	12.9	15.4	12.7
3	12.5	12.8	12.7	13.1
4	12.4	12.6	13.1	13.0
5	12.0	12.3	13.0	12.8
10	11.4	11.4	13.0	11.7
15	11.2	11.2	12.2	11.4
20	11.4	11.3	12.7	11.2

表 3 に話者適応化音声認識実験結果（単語誤り率）を示す。表から、GFDNN の場合は、適応文数が少ない場合（5 文）の場合でも適応前（適応文数 0 文）に比べて認識精度が改善できることが分かる。また、適応文数が増えると、適応パラメータ数が多い ExpFDNN でも認識精度が改善されることが分かる。CSJ は大規模なデータベースであるため、不特定話者モデル（適応化前モデル＝適応文数 0 文のモデル）でも話者の多様性に対して頑健であると考えられるが、提案手法を用いることで比較的少ない適応文数であっても認識精度の改善が得られており、提案法の有効性が示された。

今後の課題として、より複雑なネットワークである、CNN、再帰型ニューラルネットワーク（RNN)、Long Short-Term Memory（LSTM）との組み合わせについて、更なる検討を行うことが挙げられる。また、本研究では、実環境音声認識での評価として、話者適応化による評価を行ったが、雑音環境における環境適応の評価を行うことも必要であると考えられる。

＜引用文献＞

① G. Hinton, et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” IEEE Signal Processing Magazine, 29(6), pp.82-97, 2012.
② X. Feng, et al., “Speech Feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition,” Proc. ICASSP 2014, pp.1778-1782, 2014.
③ Y. Zhang, et al., “Extracting deep neural network bottleneck features using low-rank matrix facotization,” Proc. ICASSP 2014, pp.185-189, 2014.
④ T. Sainath, et al., “Deep convolutional neural networks for LVCSR,” Proc. ICASSP 2013, pp.8614-8618, 2013.
⑤ A. Biem, et al., “An application of discriminative feature extraction to filter-bank-based speech recognition,” IEEE Transactions on Speech and Audio Processing, 9(2), pp. 96-110, 2001.
⑥ F. Seide, et al., “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” Proc. ASRU, pp. 24-29, 2011.
⑦ T.N. Sainath, et al., “Learning filter banks within a deep neural network framework,” Proc. ASRU, pp. 297-302, 2013.

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 関博史，榎並大介，朱発強，山本一公，中川聖一，“話者クラスタリングに基づく短時間発話音声認識，” 電子情報通信学会論文誌，Vol.J100-D, No.1, pp.81-92, 2017, 査読有り
DOI: 10.14923/transinfj.2016JDP7063
② 藤堂祐樹，西村良太，山本一公，中川聖一，“複数の対話エージェントを用いた雑談指向の音声対話システム，” 電子情報通信学会論文誌，Vol.J99-D, No.2, pp.188-200, 2016, 査読有り
DOI:10.14923/transinfj.2015JDP7010

〔学会発表〕（計 20 件）

① Kazumasa Yamamoto, Chikara Ishikawa, Koya Sahashi, Seiichi Nakagawa, “Detection of overlapping acoustic events based on NMF with shared basis vectors,” IEEE 6th Global Conference on Consumer Electronics (GCCE 2017)（国際学会）, 2017.

② Shoko Tsujimura, Kazumasa Yamamoto, Seiichi Nakagawa, “Automatic explanation spot estimation method targeted at text and figures in lecture slides,” INTERSPEECH 2017（国際学会）, 2017.

③ Koya Sahashi, Norioki Goto, Hiroshi Seki, Kazumasa Yamamoto, Tomoyoshi Akiba, Seiichi Nakagawa, “Robust lecture speech translation for speech misrecognition and its rescoring effect from multiple candidates,” 4th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA 2017)（国際学会）, 2017.

④ Hiroshi Seki, Kazumasa Yamamoto, Seiichi Nakagawa, “A deep neural network integrated with filterbank learning for speech recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)（国際学会）, 2017.

⑤ Dairoku Kawai, Kazumasa Yamamoto, Seiichi Nakagawa, “Lyric recognition in monophonic singing using pitch-dependent DNN,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)（国際学会）, 2017.

⑥ Masashi Takebe, Kazumasa Yamamoto, Seiichi Nakagawa, “Investigation of glottal features and annotation procedure for speech emotion recognition,” 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2016)（国際学会）, 2016.

⑦ Yuma Shibahara, Kazumasa Yamamoto, Seiichi Nakagawa, “Effect of sympathetic relation and unsympathetic relation in multi-agent spoken dialogue system,” 3rd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA 2016)（国際学会）, 2016.

⑧ Dairoku Kawai, Kazumasa Yamamoto, Seiichi Nakagawa, “Speech analysis of sung-speech and lyric recognition in monophonic singing,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)（国際学会）, 2016.

⑨ Naoaki Hashimoto, Kazumasa Yamamoto, Seiichi Nakagawa, “Speech recognition based on Itakura-Saito divergence and dynamics / sparseness constraints from mixed sound of speech and music by non-negative matrix factorization,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2015)（国際学会）, 2015.

⑩ Hiroshi Seki, Kazumasa Yamamoto, Seiichi Nakagawa, “Deep neural network based acoustic model using speaker-class information for short time utterance,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2015)（国際学会）, 2015.

⑪ Akihiro Abe, Kazumasa Yamamoto, Seiichi Nakagawa, “Robust speech recognition using DNN-HMM acoustic model combining noise-aware training with spectral subtraction,” INTERSPEECH 2015（国際学会）, 2015.

６．研究組織

(1)研究代表者

山本 一公（YAMAMOTO, Kazumasa）
中部大学・工学部・准教授
研究者番号：40324230

(2)研究分担者

中川 聖一（NAKAGAWA, Seiichi）
豊橋技術科学大学・リーディング大学院教育推進機構・特命教授
研究者番号： 20115893