様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １８ 日現在

機関番号：１３９０４
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００２０１

研究課題名（和文）長時間分析窓を用いて得られる位相特徴による音声認識性能の改善に関する研究

研究課題名（英文）Improvement of speech recognition performance by using phase information with long analysis window

研究代表者

山本　一公（Yamamoto, Kazumasa）

豊橋技術科学大学・工学（系）研究科（研究院）・准教授

研究者番号：４０３２４２３０

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：これまでの音声認識技術において、音声の特徴量には、主として振幅スペクトルに基づく特徴量が用いられており、位相スペクトルに基づく特徴は無視されてきた。我々は、従来用いられてきた25ms程度の分析窓長よりも長い100～200ms程度の分析窓を用いて求めた群遅延スペクトルを特徴量として音声認識を行うことで、振幅スペクトルに基づく特徴量と同程度に音声認識が可能であることを示し、また両特徴量を併用することで音声認識精度を改善できることを示した。同時に、位相スペクトルに基づく特徴量を有効に活用するために、深層学習に基づく雑音環境下音声認識のための音響モデルに関する研究を行った。

研究成果の概要（英文）：In traditional speech recognition techniques, amplitude spectrum based features (typically MFCC or PLP) are usually used as acoustic features, while phase spectrum based features are almost ignored. In this research, we showed that the phase spectrum based features, which extracted as group delay spectrum based cepstrum features by using the longer (100-200ms) analysis window then usual one (25ms), can be used for speech recognition as the same as the amplitude spectrum based features and we can improve speech recognition performance by using the both features simultaneously. We also studied about deep learning based acoustic models for robust speech recognition in this research. We modified “noise aware training” method of Deep Neural Network based HMM (DNN-HMM) so that the DNN can treat “enhanced” noisy speech features and noise estimates. We then showed the improvement of noisy speech recognition by using the proposed method.

研究分野：音声言語情報処理

キーワード：音声認識　音響モデル　音響特徴量　位相スペクトル　群遅延スペクトル　分析窓　雑音環境　ディープニューラルネットワーク

１．研究開始当初の背景

音声認識技術は，多くの人が訓練なしに自由に使えるユーザインタフェースとして，その実用化が望まれており，特にディジタルディバイド解消の手段として期待されている．近年の研究により，静かな環境で丁寧に発声された音声に対しては，90～95%程度の単語正解精度が得られるようになっているが，背景に雑音がある環境で自由に発声した場合の単語正解精度は，70%程度に止まっており，実用化のためには性能の向上が必要不可欠となっている．

現在の音声認識システムでは，音響特徴量として MFCC（メル周波数ケプストラム係数; Mel-frequency cepstral coefficients）のような短時間フーリエ解析による振幅スペクトルに基づいた特徴量が主として用いられている．一方で，位相スペクトル特徴は，人間の聴覚が位相に鈍感であるという聴覚心理実験の結果から，音声認識のための音響特徴量としてほとんど使用されていない．いくつかある試みの中では，Murthy ら[1]の“群遅延スペクトル”に基づく特徴量がある．また，我々は，位相情報を話者認識に適用する研究を行っており，700Hz 以下の帯域の短時間位相スペクトルを用いることで，話者認識性能を向上させている．これらの結果から分かるように，短時間分析窓による位相スペクトルに音声認識に有用な情報が含まれていることは確かである．

他方，音声知覚の分野において，人間の音声知覚・認識における位相情報の重要性を報告している研究がある．Liu ら[2]は，振幅スペクトル・位相スペクトルの知覚に対する貢献度が分析窓長によって変化することを示しており，フーリエ分析窓が長くなるにつれて，その主たる知覚要因が振幅スペクトルから位相スペクトルへと変わっていくことが示されている．通常の音声特徴抽出は，上記の群遅延スペクトルを用いる方法も含めて，短時間分析を基本としており，分析フレーム長としては 25ms 程度の長さが一般的に用いられている．これに対して，上記知覚実験の結果から着想し，我々はより長い分析窓（128ms 程度）を用いて求めた位相スペクトル特徴を用いる音声認識手法を提案した．この結果として，位相スペクトル特徴単独で音声認識が可能であることが分かってきた[3].

２．研究の目的

⑴ 我々の研究[3]で，位相スペクトル情報が音声認識に有用な情報になり得ることは示した．しかし，位相スペクトル情報がどのようにして音韻的な情報を保持し，それがどのような形で音声認識精度の向上に役立っているかは明確になっていない．

本研究では，これまでの研究を発展させる形で，まず位相スペクトル特徴がどのように音声認識に役立つ情報（音韻的な情報）を保持しているのかを解明することを目的とした．これによって，どのように位相スペクトル特徴を用いることで音声認識精度が向上するかが分かるので，これまでの研究以上に位相スペクトル情報を有効に用いた音声認識が可能となる．

⑵ これまでの我々の研究から，位相スペクトル特徴は長い分析窓を用いて求めた方が，より音声認識に適した情報が得られることが分かっている．しかし，長い分析窓を用いるということは時間分解能の低下を招き，分析窓内に複数の音素が入る結果に繋がる．この場合，当該音素の前後の音素文脈まで考慮したモデル化が必要であるが，長い文脈を考慮するとモデル数が爆発的に増加し，現実的でなくなる．そのため，位相スペクトル特徴を有効に利用するための長い分析窓を使うのに適した音響モデルを開発する．

３．研究の方法

⑴ 長時間窓分析による位相スペクトル特徴量として，群遅延スペクトルを離散コサイン変換した群遅延ケプストラム特徴を用いる．

群遅延 $G(\omega)$ は以下のように定義される．

$$G(\omega) = -\frac{d\theta(\omega)}{d\omega}$$

ω は角周波数，$\theta(\omega)$ は角周波数 ω における位相である．この式を解析的に求めると，

$$G(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{\left|X(\omega)\right|^2}$$

と表すことができる．ここで，$X_R(\omega)$ および $X_I(\omega)$ は時間領域音声信号 $x(t)$ のフーリエ変換の実部・虚部をそれぞれ表し，$Y_R(\omega)$ および $Y_I(\omega)$ は，$tx(t)$ のフーリエ変換の実部・虚部をそれぞれ表す．このようにして求めた群遅延スペクトルに対して，MFCC のように離散コサイン変換を施すことで，群遅延ケプストラムを得る．位相そのものは分析窓と信号の時間位置関係によって多様に変化するが，群遅延スペクトルは分析窓の時間位置変化に対して比較的頑健であるため，これを位相スペクトル特徴量として用いる．

従来の音声認識システムでは，特徴抽出用の分析窓として，20～30ms 程度の短時間分析窓が使用されている．これは，この程度の長さで音声信号を見ると，準定常信号とみなすことができるためである．しかし，我々の従来の研究[3]により，短時間分析窓の代わりに，より長い時間分析窓（100～200ms 程度のもの）を用いることで，安定した位相スペクトル（群遅延スペクトル）が求まることが分かっている．図 1 および図 2 に具体例を示す．図 1 は短時間分析窓（25ms）を用いて分析した群遅延スペクトル，図 2 は長時間分析窓（128ms）を用いて分析した群遅延スペクトルである．この図からも分かるように，長時間窓を用いて分析を行うことで，スペク

トルの概形がはっきりと分かるようになっている．このため，本研究においても長時間分析窓を用いて求めた群遅延スペクトル（実際には，群遅延ケプストラム）を位相特徴として用いる．

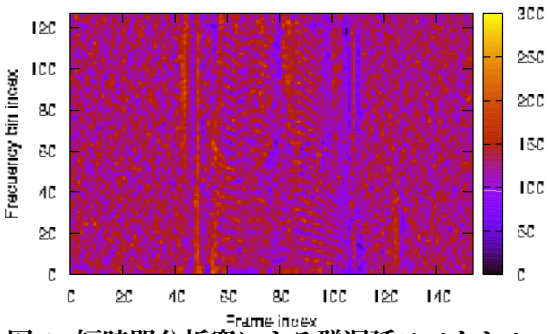

図 1. 短時間分析窓による群遅延スペクトル

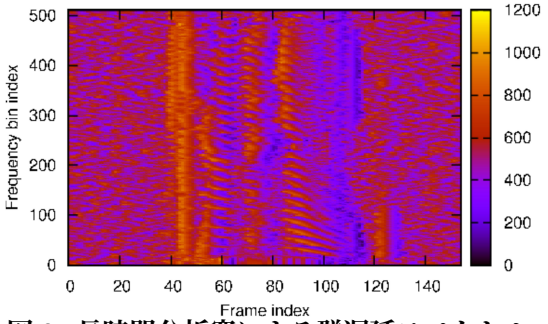

図 2. 長時間分析窓による群遅延スペクトル

(2) 長時間分析窓による位相スペクトル特徴を用いる際に問題となるのは，時間分解能の低下である．分析フレームのシフト幅を従来の短時間分析窓での分析と同様（10ms 程度）とすることで，フレーム数は同じにできるが，比較的似たような特徴量がベクトル時系列として続くことになり，これらから上手く情報を抽出するモデルが必要となってくる．

これに対して本研究では，最近音声認識でも活発に研究されるようになった，深層学習（ディープラーニング）を用いる．具体的には，複数の隠れ層を持つニューラルネットワークであるディープニューラルネットワーク（Deep Neural Network; DNN）を用いる．DNN の例を図 3.に示す．DNN は入力層，複数の隠れ層，出力層によって構成される．

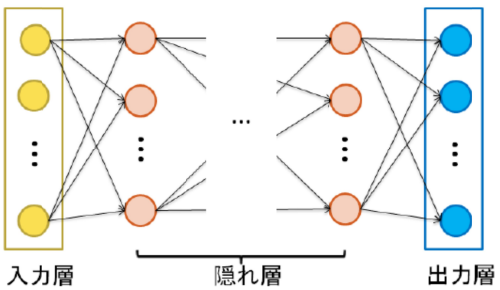

図 3. ディープニューラルネットワーク

これまでの音声認識技術では，音響特徴ベクトルの時系列に対して，その変動を吸収するための確率的な手法として，隠れマルコフモデル（Hidden Markov Model; HMM）が広く用いられてきた．図 4 に HMM の概形を示す．

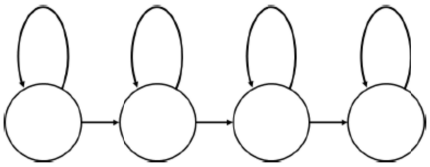

図 4. 隠れマルコフモデルの概形

通常の HMM では，特徴ベクトルの出力確率を求めるために，ガウス混合分布（Gaussian Mixture Model; GMM）が用いられている．これに対して HMM と DNN を組み合わせて用いる DNN-HMM においては，DNN の出力層のユニットは HMM の出力状態に対応しており，以下の式を計算することで，GMM によって計算されていた状態 z_t における特徴ベクトル x_t の出力確率 $p(x_t \mid z_t)$ を，DNN によって計算する．

$$p(x_t \mid z_t) = \frac{p(z_t \mid x_t)}{p(z_t)} p(x_t)$$

DNN は状態の事後確率 $p(z_t \mid x_t)$ を学習する．$p(x_t)$ は音声認識を行う際に定数として扱うため無視され，$p(z_t)$ は学習データをカウントすることで求めることができる．DNN の学習は教師データを用いたバックプロパゲーション法によって行う．DNN はネットワークを深く（多層に）することで複雑なモデルを構築できるが，ランダムな初期値では過学習が容易に起こり，上手く学習を行うことが困難であった．そこで，Hinton らは制約付きボルツマンマシン（Restricted Boltzmann Machine; RBM）を複数つなげた Deep Belief Network （DBN）を初期値とする手法を提案した[4]．この手法により層が深い場合でも過学習を避けることが可能となった．DNN は GMM と比較するとより大きい次元数の特徴パラメータを扱えるなどの特徴がある．一般的に，GMM-HMM では短時間の特徴パラメータ 1 フレームを入力とするが，DNN－HMM では短時間の特徴パラメータを前後 5 フレーム含め計 11 フレームを入力とする．これは，より大きい次元数を扱っていると同時に，より長時間の情報を扱っていると言える．例えば，MFCC は一般的にフレーム長 25ms，フレームシフト 10ms で特徴抽出を行うため，11 フレームを合わせると 125ms 分の音声信号の情報を用いていることになる．一方，長時間分析窓に基づく群遅延ケプストラムはフレーム長が 128ms と，MFCC 11 フレーム分の音声情報をカバーしている．この性質により，予備実験において GMM-HMM では MFCC と長時間分析窓に基づく群遅延ケプストラムを併用した場合の音素認識精度が MFCC 単独の場合とほぼ変わらなかったが，DNN-HMM では改善

が見られるのではないかと考えた．

また，雑音環境下において DNN-HMM を頑健に動作させるために，通常は様々な雑音環境下の音声を学習データとして用いて，雑音環境に対して DNN を汎化する，Multi-condition training が行われる．これに対して，DNN の入力層に現在の雑音環境の情報（推定した雑音スペクトルの情報）を与えることで，更に耐雑音性が向上するという，“noise-aware training”という手法が提案されている[5]．雑音抑圧を行った音響特徴量を用いた学習（noise adaptive training）と呼ばれる手法も一般的に行われている．我々は，これらを組み合わせることで，雑音抑圧を行った音響特徴量に対する信頼性を雑音環境情報により与えられ，それによって雑音環境下での音声認識性能を改善することができるのではないかと考えた．

４．研究成果

(1) 位相スペクトル特徴である長時間分析群遅延ケプストラム係数と，従来法である MFCC を組み合わせて DNN-HMM による認識実験を行った．

認識対象として，日本語連続数字データベースである CENSREC-1 を使用した．学習データとして，CENSREC-1 の clean training data と multi-condition training data を，テストデータとして set B のものを用いた．

音響モデルは DNN-HMM で，モデル単位として音素（monophone）を使用している．各音素モデルの状態数は 3 である．DNN の入力は，“MFCC”として現在のフレームに前後 5 フレームずつを結合した 11 フレーム分の MFCC 特徴ベクトルを結合したもの（Δ特徴を含む 39 次元×11 フレーム）を用い，“位相”としては現在のフレームのみ（1 フレーム分）の長時間分析群遅延ケプストラム（20 次元）を用いた．DNN の隠れ層は 5 層で，各隠れ層は 512 ノードとした．

表 1 に clean training data で学習した場合の実験結果を示す．数値は数字正解精度である．

表 1. clean training の結果

Test SNR	MFCC のみ	MFCC+位相
Clean	75.21	79.11
20dB	69.73	69.20
15dB	67.07	66.61
10dB	60.33	59.12
5dB	47.83	46.11
0dB	26.58	12.97
-5dB	8.13	6.00

結果から，位相特徴を加えることで，テストデータが clean の条件では認識精度が改善しているが，雑音が加わると MFCC のみの場合に比べて認識精度が低下する結果となっていることが分かる．

表 2 に multi-condition training data で学習した場合の実験結果を示す．

表 2. Multi-condition training の結果

Test SNR	MFCC のみ	MFCC+位相
Clean	86.21	90.05
20dB	76.61	77.17
15dB	74.51	75.22
10dB	69.27	69.12
5dB	58.23	57.08
0dB	33.78	30.49
-5dB	8.40	5.44

こちらは，MFCC のみの場合と比べて，位相特徴を加えることで，テストデータの SNR が clean～15dB では認識精度が改善できていることが分かる．これらの結果から，位相スペクトル特徴は雑音の影響を受けやすく，SNR が低い環境においては，特徴量としてそのまま用いるのが難しいのではないかと考えられ，当初の期待に反する結果となった．

クリーン音声を対象として大語彙連続音声認識システムでも同様の実験を試みたが，大語彙連続音声認識の場合は，クリーン音声でも，位相情報の追加による認識率の改善はほとんど見られなかった．

位相スペクトルは，分析フレーム内における信号の時間配置（分析フレーム内のエネルギー重心）と関係があり，中村らは群遅延スペクトルと深い関係のある平均時間スペクトルを帯域別に求めることで，群遅延スペクトルと同じような特性の特徴量を安定して求めることができ，それによって雑音環境下で音声認識性能が改善できることを示している[6]．図 1 および図 2 に示したように，群遅延スペクトルは，振幅スペクトルとも非常に近い関係にあるため，これらのことから，位相特徴量を用いることで，フレーム単位で分析（音響特徴抽出）を行うことで失われてしまうフレーム内の時間情報をある程度補うことができており，それによって認識精度の改善が得られていると考えられる．しかし，雑音環境下においては雑音の影響により時間情報を正確に表現することができず，また連続音声に対しては分析時間が長くなることで時間分解能の低下が大きな問題となり，それぞれの環境で認識精度の向上が得られないのだと考える．以上のことから，今後はガンマトーンフィルタバンクを通して得た時間領域信号のエネルギー包絡を活用するなどして，雑音環境下や連続音声においても時間情報を上手く扱う手法について取り組んでいく予定である．

(2) DNN-HMM の学習に対して，noise-aware training と noise adaptive training を融合した手法を提案する．

雑音抑圧手法としては，単純なスペクトルサブトラクション（Spectral Subtraction; SS）法を用いた．雑音スペクトルの推定手法としては，発話開始前のいくつかのフレームを雑

音とみなして平均スペクトルを求め，その発話中は常に同じ雑音スペクトル情報を使い続ける方法（utter.）と，フレーム毎に Minimum Statistics（MS）に基づいてフィルタバンク出力のエネルギーの底を推定することで，アンビエント雑音を推定するもの（frame）の 2 種類を行った．

認識対象としては，先ほどの実験と同じく日本語連続数字データベースである CENSREC-1 を使用した．DNN-HMM に入力する特徴量としては，MFCC(Δ 特徴を加えた 39 次元)の他に，MFCC に変換する前のメルフィルタバンク出力（FBANK．フィルタ数は 30），および，ガンマトーンフィルタバンク出力（GFBANK．フィルタ数は 30）を用いた．先ほどの実験と同様に，各特徴量を 11 フレーム結合したものを DNN の入力として用いている．noise-aware training および noise-aware training と noise adaptive training を組み合わせる提案法に用いる雑音スペクトル情報は，1 フレームのみとして，音響特徴量と同じ処理によって抽出した特徴量を用いた．位相情報が雑音に強くないことが分かったため，この実験では位相情報を用いていない．

図 5 に noise-aware training と noise adaptive training を組み合わせる提案手法（図中の“proposed”）と，各従来法の比較を示す．グラフの縦軸は数字正解精度である．DNN の隠れ層が 3 層のものから 8 層のもので実験を行い，もっとも精度が良かったものをグラフで示している．

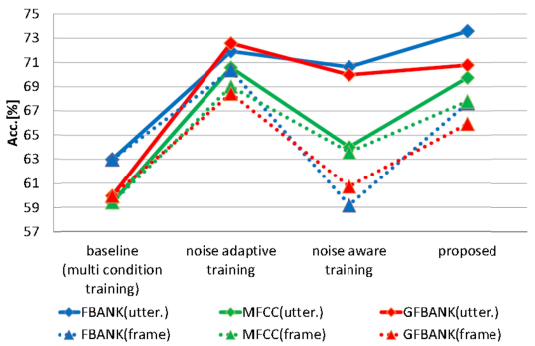

図 5. noise-aware training と noise adaptive training の組み合わせによる効果

特徴量ごとに各手法の比較を行うと，発話ごとの雑音推定において，FBANK では提案手法，GFBANK と MFCC では雑音抑圧のみを行った場合が最も精度が良かった．特徴量によって傾向が変わった理由として，フィルタバンクチャネル出力に対する値の圧縮方法の差が挙げられる（MFCC と FBANK は対数圧縮，GFBANK はルート圧縮）．この違いにより，雑音抑圧後の特徴量と推定雑音情報間の学習において差が生じたと考えられる．ルート圧縮を行った特徴が DNN と相性が悪く，ルート圧縮を行った特徴量と推定雑音情報間の関係性を十分に学習するには深いネットワークが必要であり，そのためにより多くの学習データが必要なのではないかと推測している．

また，雑音推定単位（発話単位“utter.”，フレーム単位“frame”）の間で認識精度を比較すると，どの特徴量，手法においても，フレームごとの推定よりも発話ごとの推定の方が精度が良くなった．発話ごとの推定では発話前のフレームを用いているが，その中のほとんどのフレームが純粋な雑音区間であることから，定常な雑音を上手く推定できていると考えられる．一方で MS を用いたフレームごとの推定では，特に低 SNR のときに音声と雑音の区別が困難であるために雑音推定が上手くいかず，その上手く推定出来なかった推定雑音情報を直接 DNN-HMM に入力として与えたことで認識精度が悪くなったと考えられる．つまり，noise-aware training や本研究の提案手法は雑音抑圧のみを行う場合よりも雑音推定精度に依存する傾向があると言える．

今後の課題として，今回は一般的な特徴量を用いたが，Power Normalized Spectrum（PNS）のような最新のノイズロバストな特徴量に対して今回の提案手法を適用することが挙げられる．また，より高精度なフレームごとのノイズ推定方法の検討も必要である．

今後は，位相特徴に関する研究の知見と，DNN の雑音頑健性向上に関する知見を組み合わせることで，実環境下でより頑健に動作する音声認識システムの開発を目指す。

＜引用文献＞

① H. A. Murthy, V. Gadde, “The modified group delay function and its application to phoneme recognition,” ICASSP 2003, pp.I-68–I-71, 2003.

② L. Liu, J. He, G. Palm, “Effects of phase on the perception of intervocalic stop consonants,” Speech Communication, Vol.22, No.4, pp.403–417, 1997.

③ 山本，中川，“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, 2011.

④ G. E. Hinton, S. Osindero, and Y. The, “A fast learning algorithm for deep belief nets,” Neural Computation, 18:1527 1554, 2006.

⑤ M. L. Seltzer, D. Yu and Y. Wang. “An Investigation of Deep Neural Networks for Noise Robust Speech Recognition”, ICASSP 2013, pp.7398-7402, 2013.

⑥ 中村，藤村，篠原，益子，河村，“群遅延に基づく音声特徴量の雑音環境下での評価,” 日本音響学会 2012 年春季研究発表会講演論文集, 2012.

５．主な発表論文等

〔雑誌論文〕（計　1　件）

① A. A. Nugraha, K. Yamamoto, S. Nakagawa, “Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition,” EURASIP Journal on Audio, Speech, and Music Processing, 2014:13, Jan. 2014. **（査読有）** DOI:10.1186/1687-4722-2014-13

〔学会発表〕（計　14　件）

① A. Abe, K. Yamamoto, S. Nakagawa, “Robust Speech Recognition using DNN-HMM Acoustic Model Combining Noise-aware training with Spectral Subtraction,” INTERSPEECH2015, Dresden, Germany, Sep. 2015. **(採択決定)（査読有）**

② **橋本尚亮**, **山本一公**, **中川聖一**, “NMF **と雑音適応学習に基づく音楽重畳音声認識に関する検討**,” **日本音響学会** 2015 **年春季講演論文集**, 1-P-11, 2015.

③ **阿部晃大**, **山本一公**, **中川聖一**, “Noise-aware training **と** SS **を併用した** DNN-HMM **音響モデルの雑音下音声認識の評価**,” **日本音響学会** 2015 **年春季講演論文集**, 1-P-14, 2015.

④ **関博史**, **山本一公**, **中川聖一**, “**年齢性別クラスタリング情報を考慮した** DNN-HMM **による音声認識の検討**,” **日本音響学会** 2015 **年春季講演論文集**, 1-P-28, 2015.

⑤ N. Hashimoto, K. Yamamoto, S. Nakagawa, “Speech recognition based on Itakura-Saito divergence and dynamics / sparseness constraints from mixed sound of speech and music by non-negative matrix factorization,” INTERSPEECH 2014, pp.2749-2753, Singapore, Sep. 2014. **（査読有）**

⑥ H. Seki, K. Yamamoto, S. Nakagawa, “Comparison of syllable-based and phoneme-based DNN-HMM in Japanese speech recognition,” ICAICTA 2014, pp.249-254, Bandung, Indonesia, Aug. 2014. **（査読有）**

⑦ A. A. Nugraha, K. Yamamoto, S. Nakagawa, “Single channel dereverberation method in log-Mel spectral domain using limited stereo data for distant speaker identification,” APSIPA ASC 2013, CD-ROM, Kaohsiung, Taiwan, Oct. 2013. **（査読有）**

⑧ S. Nakano, K. Yamamoto, S. Nakagawa, “Fast NMF based approach and VQ based approach using MFCC distance measure for speech recognition from mixed sound,” APSIPA ASC 2013, CD-ROM, Kaohsiung, Taiwan, Oct. 2013. **（査読有）**

⑨ **橋本尚亮**, **仲野翔一**, **山本一公**, **中川聖一**, “NMF **による音楽重畳音声の音声認識の改善**,” **日本音響学会** 2013 **年秋季講演論文集**, 1-P-2b, 2013.

⑩ J. McDonough, K. Kumatani, T. Arakawa, K. Yamamoto, B. Raj, “Speaker tracking with spherical microphone arrays,” ICASSP 2013, pp.3981-3985, Vancouver, Canada, May 2013. **（査読有）**

⑪ **仲野翔一**, **山本一公**, **中川聖一**, “**ケプストラム距離に基づく** NMF **の高速化手法と** VQ **手法による音楽重畳音声の認識**,” **日本音響学会** 2013 **年春季講演論文集**, 1-Q-26b, 2013.

⑫ K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, I. Tashev, “Microphone Array Processing for Distant Speech Recognition: Towards Real-World Deployment,” APSIPA ASC 2012, Hollywood, USA, Dec. 2012.**（査読有）**

⑬ D. Enami, F. Zhu, K. Yamamoto, S. Nakagawa, “Soft-clustering Technique for Training Data in Age- and Gender-independent Speech Recognition,” APSIPA ASC 2012, Hollywood, USA, Dec. 2012.**（査読有）**

⑭ S. Nakano, K. Yamamoto, S. Nakagawa, “Fast NMF Based Approach and Improved VQ Based Approach for Speech Recognition from Mixed Sound,” APSIPA ASC 2012, Hollywood, USA, Dec. 2012.**（査読有）**

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・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・特任教授

研究者番号：２０１１５８９３