Vol.2017-SLP-115 No.12
2017/2/18

ChimeChallengeタスクにおける NMFによる雑音除去の検討

小澤 奈摘[1]　田中 智大[1]　篠崎 隆宏[1,a)]

概要：近年、情報技術の発展により音声認識システムが広く使用されるようになっている。現在の音声認識システムの問題点として、雑音環境下での認識率の低下があげられる。特に課題となっているのが、シングルマイクでの効果的な雑音除去法の実現である。本研究では、バスやカフェ内などの雑音が存在する環境下で音声認識機能を備えたタブレットを使用することを想定した、Chime Challenge タスクの音声データを対象とした音声認識実験に取り組んだ。シングルマイクによって録音された雑音重畳音声を対象に、非負値行列因子分解法を応用した雑音除去法により認識率の向上を狙い実験を行った。単純に NMF を組み込むだけでは認識率は向上しなかったが、雑音を取り除いた音源と雑音重畳音声の特徴量を連結することにより認識率が 1%程度向上することを示した。

キーワード：シングルチャネル音源分離、ChimeChallenge、非負値行列因子分解

1. はじめに

近年、情報技術の発展により身の回りの様々なものにコンピュータが使われ、スマートフォンやタブレット端末などの電子機器がより身近になっている。これらの端末には様々な技術が搭載されており、その中でも音声認識技術は音声検索、カーナビゲーションやゲームなどの分野に使われ、屋内だけでなくあらゆる場面で利用できるようになった。しかし、このような音声認識技術における課題として、雑音環境下での認識精度の低下が挙げられる。そのため、快適に音声認識技術を利用するには雑音を低減させることが重要となる。

雑音を低減させる手法の一つとしてシングルチャネル音源分離がある。これは、一本のマイクロホンのみを用いて音声と雑音を分離する手法であり、マイクの数が一本で済むことから複数のマイクロホンを用いるマルチチャネル音源分離よりも便利であるという特徴がある。しかし、分離をする際に音源の位置情報を利用することができないため難度の高い技術となっている。

シングルチャネル音源分離を行う手法として非負値行列因子分解 (NMF : Nonnegative Matrix Factorization) が提案されている。これは音声のスペクトログラムを行列とみなすことにより音声データを 2 つの行列の積によって近似する手法である。これにより音声データをコンパクトな空間で表すことができ、それをもとに音声と雑音を分離をすることが可能となる。

耐雑音音声認識性能の向上を目的とした国際的なワークショップとして ChimeChallenge [*1]がある。これは主催者より提供された音声認識タスクに対して、様々な参加者がそれぞれの音声認識システムの認識性能を競うものである。認識タスクは毎年更新されるが、使用するマイクロホンの数などに応じて複数のカテゴリに分かれている。また、認識タスクとともに音声認識ツールキット Kaldi を用いたベースライン認識システムのレシピも用意されている。ベースラインレシピには前年までの ChimeChallenge ワークショップの成果が反映されており、これを用いることで高性能なベースラインシステムを用いた音声認識実験を容易に行うことができる。本研究ではシングルマイクロホンを用いたタスクを対象とし、ChimeChallenge より提供されているベースラインに NMF を組み込むことで認識性能の向上を図る。

本論文の構成は以下に示す通りである。第 2 章では音源分離手法として NMF について説明する。第 3 章では ChimeChallenge タスクについて説明する。第 4 章で本研究で構築した NMF を用いたシステムについて説明し、第 5 章で実験条件と実験結果を示す。最後に第 6 章で本論文のまとめと今後の課題について述べる。

[1] 東京工業大学
Tokyo Institute of Technology
[a)] www.ts.ip.titech.ac.jp

[*1] http://spandh.dcs.shef.ac.uk/chime_challenge/

2. 非負値行列因子分解（NMF）

2.1 NMF について

NMF とは非負値のみからなる行列を分解する手法である [1][2][3]。文書や画像、音声などのデータは非負値によって表すことができる。文書データの場合は文章を構成している単語数、画像の場合は画素値などの画像データ、そして音声の場合はパワースペクトルなどを利用してそれぞれの構成成分を抽出している。NMF はこのような非負値データを表現した行列を分解することができるため幅広い分野に応用することができる。

もともとは画像処理分野で生まれた技術であり、顔画像から顔パーツを抽出する目的で NMF が使われた [4]。現在、音のスペクトルを行列として適用し NMF を使用することによって、音源分離だけでなく自動採譜 [5] など様々な音声分野においても重要な技術となっている。

非負値行列因子分解は上記に挙げたような行列の要素が 0 か正の値となる信号やデータを対象としている。NMF による分解結果として得られるものは、それぞれのデータに対するいくつかの頻出するパターンである。これにより幅広い分野のデータを解析することが可能となっている。

2.2 NMF アルゴリズム

NMF では式 (1) に示すように、行列 $\boldsymbol{V}$ を 2 つの非負値行列の積 $\boldsymbol{W}$ と $\boldsymbol{H}$ の積に分解する。

$$\boldsymbol{V} \simeq \boldsymbol{W}\boldsymbol{H} \tag{1}$$

このように行列を 2 つの非負値行列の積によって近似する考え方は Paatero や Lee によって提案されている [1][2]。分解方法としては行列 $\boldsymbol{V}$ と行列 $\boldsymbol{W}$、$\boldsymbol{H}$ との距離を最小にすることを考える。この距離を計算する際、二乗誤差、I ダイバージェンス、板倉斎藤距離 [6] の 3 種類が主に使われている。

2.3 NMF を利用した音源分離

音のスペクトログラムを行列とみなすことで、音声データも NMF により非負値行列の積に分解できる。これにより音源分離や雑音除去が可能となる [7][8][9]。雑音の混ざった音声データから雑音のないクリーンな音声のみを取り出す手法を以下で説明する。

音のパワースペクトルを行列 $\boldsymbol{V}$ とした時、行列 $\boldsymbol{W}$ と $\boldsymbol{H}$ に分解することを考える。ここで、行列 $\boldsymbol{V}$ のサイズを $d \times T$ としたとき、行列 $\boldsymbol{W}$、$\boldsymbol{H}$ のサイズはそれぞれ $d \times k$、$k \times T$ と表すことができる。ここで、d は周波数ビン数、T は時間フレーム数、k は基底数となる。行列 $\boldsymbol{W}$ の各列は音声の頻出パターンを表現し、行列 $\boldsymbol{H}$ はそれらの線形和を作る際の重みを表現する。また、基底数 k によって音声データの表現に用いる頻出パターンの数が決まる。雑音除去ではまず学習段階において、行列によるスペクトルの表現とその分解をクリーン音声と雑音のそれぞれについて行う。学習用のクリーン音声のパワースペクトルを表す行列を $\boldsymbol{V_s}$ とし、それを NMF によって分解した行列を式 (2) に示すように $\boldsymbol{W_s}$、$\boldsymbol{H_s}$ とする。同様に、雑音のみのパワースペクトルを表す行列を $\boldsymbol{V_n}$ とし、それを NMF によって分解した行列を式 (3) に示すように $\boldsymbol{W_n}$、$\boldsymbol{H_n}$ とする。

$$\boldsymbol{V_s} \simeq \boldsymbol{W_s}\boldsymbol{H_s} \tag{2}$$

$$\boldsymbol{V_n} \simeq \boldsymbol{W_n}\boldsymbol{H_n} \tag{3}$$

分解により得られたそれぞれの頻出パターンを表す行列 $\boldsymbol{W_s}$ と $\boldsymbol{W_n}$ を式 (4) に示すように連結した行列を $\boldsymbol{W}$ とする。

$$\boldsymbol{W} = [\boldsymbol{W_s} \ \ \boldsymbol{W_n}] \tag{4}$$

音声認識時における雑音除去では、この行列 $\boldsymbol{W}$ を固定した状態で式 (5) に示すように入力雑音重畳音声のスペクトログラム $\boldsymbol{V}$ に対し NMF を適用する。重みを表す行列 $\boldsymbol{H}$ は，適当な初期値をもとに反復法により入力に対応した値を求める。

$$\boldsymbol{V} \simeq \boldsymbol{W}\boldsymbol{H} \tag{5}$$

その後、行列 $\boldsymbol{W}$ と $\boldsymbol{H}$ からクリーン音声に対応した部分行列である $\boldsymbol{W_s}$ と $\boldsymbol{H_s}$ を取り出し再合成することにより、音声のみを取り出すことができる。

3. ChimeChallenge タスク

3.1 タスク概要

ChimeChallenge は公共エリアでの雑音による雑音環境下音声認識を対象タスクとして国際的なコンテスト形式で音声認識性能を競うもので、今までに 4 回ワークショップが開催されている。世界中の企業や研究機関などが参加し、成果を競い合っている。今回本研究で対象としたタスクは 4 回目に開催された Chime-4 チャレンジタスク (The 4th CHiME Speech Separation and Recognition Challenge) 用に公開されたものである。雑音の種類としては、バスの車内、カフェテリア、歩行エリア、道路の 4 種類の雑音があり、これらの雑音が重畳された雑音重畳音声において音声認識を行う。

4. NMF を用いた CHiME 認識システム

CHiME 認識システムにおける NMF の組み込み方について説明する。Chime-4 チャレンジタスクでは雑音重畳音声から音声認識を行い認識結果を出す流れになっている。ここで、NMF をフロントエンド部分に追加し、雑音の除去された音源に対して音声認識を行う。

Chime システムではまず特徴量とし

ては 13 次元の MFCC とその Δ、Δ Δ 計 39 次元を用いている。認識システムの構築では、まず MFCC を用いて GMM-HMM を学習する。それを元に特徴量として MFCC ベクトルを前後 3 フレームについて連結し LDA および MLLT を適用した 40 次元のベクトルを用いて GMM-HMM を学習する。さらに、fMLLR を用いた話者適応学習 (SAT) を行う。そしてこれにより得られた話者正規化特徴量とアライメントを元に、DNN-HMM の学習を行う。DNN-HMM では、特徴量として、話者正規化特徴量を前後 5 フレーム連結したものを用いる。本論文の実験では、DNN-HMM はクロスエントロピー基準で学習し、系列識別学習やリスコアリングは行っていない。言語モデルには 3-gram を用いた。

NMF による雑音除去音声を用いた認識実験では、雑音除去音声を元にした MFCC をそのまま用いたシステムと、雑音が重畳した音声から作成した MFCC と雑音除去音声を元にした MFCC を特徴量ベクトルの領域で次元方向に連結した特徴量を用いたシステムの 2 通りを検討した。

5. 実験

5.1 実験条件

実験には ChimeChallenge タスクにて用いる音声データ及び雑音データを用いた。雑音はバスの車内での雑音 (BUS)、カフェテリア内での雑音 (CAF)、歩行エリアでの雑音 (PED)、道路での雑音 (STR) の 4 種類を用いた。学習データ (tr05) として男性話者 2 名女性話者 2 名による実際の雑音環境下で収録された 1600（各雑音環境 400）発話と、男性話者女性話者合わせて 83 名の話者によるシミュレーションされた雑音が付加された 7138（BUS：1728、CAF：1794、PED：1765、STR：1851）文章の計 8738 文章を用いた。開発データ (dt05) として男性話者 2 名女性話者 2 名による実際の雑音環境下で収録された 1640（各雑音環境 410）文章と、シミュレーションされた雑音が付加された 1640（各雑音環境 400）文章を用いた。また、評価データ (et05) として男性話者 2 名女性話者 2 名による実際の雑音環境下で収録された 1320（各雑音環境 330）文章と、シミュレーションされた雑音が付加された 1320（各雑音環境 330）文章を用いた。

NMF はそれぞれの雑音ごとに作成し実験を行っている。音声データのサンプリング周波数は 16kHz である。音声から NMF 用にスペクトルを求める際の窓幅は 512、シフト幅は 256 とした。NMF の基底数は事前実験をもとに 60 とした。

5.2 実験結果

表 1 は GMM-HMM を用いた場合の認識実験の結果である。評価尺度としては単語誤り率 (WER : Word Error Rate) を用いた。ここで、real は実際の雑音重畳音声、simu はシミュレーションされた雑音が付加された音声である。雑音重畳音声から作成した MFCC をそのまま用いたベースラインを用いた場合 (baseline)、NMF を適用した音声から抽出した特徴量を使用した場合 (mmf_o)、NMF を適用して作成した MFCC と雑音重畳音声から作成した MFCC を連結した特徴量を用いた場合 (nmf_c) の順に表にまとめた。また、表 2 は同様に DNN-HMM を用いた場合の認識実験の結果である。表 1、2 どちらの場合も、特徴量を連結しないで行った実験では WER の値がベースタインよりも上昇し悪化してしまっている。これは NMF を適用した場合、雑音は期待通り抑圧されるものの音声の情報も一部失われてしまうためと考えられる。他方, 連結した特徴量を用いると WER の値が下がり認識性能が向上していることがわかる。これは特徴量の領域において雑音除去前の音声の情報を保ちつつ、雑音を抑圧した効果を加えることができたためと考えられる。

表 1 音響モデルとして GMM-HMM を用いて認識を行った場合の単語誤り率 (WER)

	dt05(real)	dt05(simu)	et05(real)	et05(simu)
baseline	22.18	24.47	37.62	33.26
nmf_o	26.59	31.90	41.54	40.11
nmf_c	21.06	24.35	36.89	32.39

表 2 音響モデルとして DNN-HMM を用いて認識を行った場合の単語誤り率 (WER)

	dt05(real)	dt05(simu)	et05(real)	et05(simu)
baseline	16.46	17.53	29.84	26.22
nmf_o	19.25	24.01	33.95	30.04
nmf_c	15.58	17.06	28.42	25.04

表 3 シミュレーションされた雑音重畳音声による開発データの単語誤り率 (WER)。音響モデルに GMM-HMM を使用

	BUS	CAF	PED	STR
baseline	20.65	29.60	20.99	26.64
nmf_o	21.35	45.51	27.50	33.27
nmf_c	20.27	30.62	19.86	26.66

表 4 実際の雑音重畳音声による開発データの単語誤り率 (WER)。音響モデルに GMM-HMM を使用

	BUS	CAF	PED	STR
baseline	27.84	22.35	16.99	21.53
nmf_o	25.58	36.93	19.44	24.41
nmf_c	25.59	22.84	16.61	20.86

表 3〜6 は GMM-HMM を用いた認識実験において各雑音ごとの WER を示した詳細であり、表 7〜10 は DNN-HMM

表 5 シミュレーションされた雑音重畳音声による評価データの単語誤り率 (WER)。音響モデルに GMM-HMM を使用

	BUS	CAF	PED	STR
baseline	26.93	38.25	33.97	33.88
nmf_o	25.43	48.41	46.20	40.38
nmf_c	25.24	39.01	34.71	30.60

表 6 実際の雑音重畳音声による評価データの単語誤り率 (WER)。音響モデルに GMM-HMM を使用

	BUS	CAF	PED	STR
baseline	52.05	40.77	33.69	23.96
nmf_o	51.42	45.71	41.01	28.05
nmf_c	50.66	41.44	32.68	23.79

表 7 シミュレーションされた雑音重畳音声による開発データの単語誤り率 (WER)。音響モデルに DNN-HMM を使用

	BUS	CAF	PED	STR
baseline	15.90	21.03	14.10	19.10
nmf_o	19.25	30.08	21.66	25.01
nmf_c	12.85	22.00	14.68	18,69

表 8 実際の雑音重畳音声による開発データの単語誤り率 (WER)。音響モデルに DNN-HMM を使用

	BUS	CAF	PED	STR
baseline	21.35	17.20	11.20	16.09
nmf_o	24.69	21.50	13.33	17.46
nmf_c	19.51	17.73	10.72	14.35

表 9 シミュレーションされた雑音重畳音声による評価データの単語誤り率 (WER)。音響モデルに DNN-HMM を使用

	BUS	CAF	PED	STR
baseline	21.12	30.39	26.17	27.19
nmf_o	24.55	35.30	29.81	30.40
nmf_c	20.33	30.41	25.18	24.22

表 10 実際の雑音重畳音声による評価データの単語誤り率 (WER)。音響モデルに DNN-HMM を使用

	BUS	CAF	PED	STR
baseline	42.91	32.33	26.35	17.80
nmf_o	47.03	37.67	31.40	19.69
nmf_c	41.24	31.37	25.64	15.42

を用いた認識実験において各雑音ごとの WER を示した詳細である。特徴量を連結していない場合、GMM-HMM を用いた時では BUS の雑音でいくつか値の向上が見られたものの、他の雑音では値が悪くなってしまっている。DNN-HMM を用いた時では全ての雑音に関して値が悪くなってしまっている。特徴量を連結した場合、GMM-HMM と DNN-HMM のどちらを用いた時でも CAF 以外の雑音では向上が見られたが、CAF の雑音に対してはほとんどの場合で値が悪くなっている。これは、CAF の雑音には人の声が多く含まれるため他の雑音に比べて NMF を適応するのが難しいという理由によるものと思われる。

6. まとめ

Chime Challenge タスクの音声データを対象とした音声認識実験に取り組み、非負値行列因子分解法を応用した雑音除去法による認識率の向上を目的に実験を行った。単純に NMF だけを組み込んだもの、雑音除去前および除去後の特徴量を連結したものそれぞれについて認識実験を行い、特徴量を連結しないものに関しては認識性能は向上しなかったが、特徴量を連結したものは認識性能が向上することを示した。雑音ごとの結果では、連結した特徴量を用いることで BUS、PED、STR の雑音に関して認識性能の向上が見られたが、CAF の雑音に関しては向上が見られなかった。今後の課題として雑音の種類によらず認識性能の安定した向上が得られるようにすることなどが挙げられる。

謝辞 本研究は JSPS 科研費 26244026、16H01935 の助成を受けたものです。

参考文献

[1] D.D.Lee and H.S.Seung, "Learning the parts of objects with nonnegative matrix factorization", Nature, 401, 788/791 (1999).
[2] P.Paatero and U.Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", Environmetrics, 5, 111/126 (1994).
[3] 澤田宏, "非負値行列因子分解 NMF の基礎とデータ/信号解析への応用", 信学誌, vol.95, no.9 (2012).
[4] D.D.Lee and H.S.Seung, "Algorithms for nonnegative matrix factorization", NIPS, 556/562 (2000).
[5] P.Smaragdis and J.C.Brown, "Non-negative matrix factorization for music transcription", Proc. WASPAA 2003, 177-180 (2003).
[6] C.F é votte, N.Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis", Neural Computation, vol. 21, no. 3, 793-830 (2009).
[7] M.N.Schmidt, J.Larsen, and F.-T.Hsiao, "Wind noise reduction using non-negative sparse coding", Machine Learning for Signal Processing, IEEE, 431-436 (2007).
[8] T.Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria", IEEE Trans. Audio, Speech, Lang. Process., Vol.15, no. 3, 1066-1074 (2007).
[9] M.N.Schmidt and R.K.Olsson, "Single-channel speech separation using sparse non-negative matrix factorization", INTERSPEECH, 2614-2617 (2006).