法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

ASAMモデルによる話者認識のための特徴量に関する研究

著者	宋 豫
出版者	法政大学大学院情報科学研究科
雑誌名	法政大学大学院紀要．情報科学研究科編
巻	15
ページ	1-6
発行年	2020-03-24
URL	http://doi.org/10.15002/00022726

ASAMモデルによる話者認識のための特徴量に関する研究
A study on features for speaker recognition by ASAM model

宋 豫 *
Song Yu
法政大学大学院　情報科学研究科　情報科学専攻
Email: yu.song.5m@stu.hosei.ac.jp

***Abstract*—For multi-speaker recognition, deep learning-based frameworks have made significant progress in multi-speaker mixed speech separation, but are unable to provide satisfactory solutions in complex auditory scenarios. A unified auditory selection framework with attention and memory can solve this problem. First, the sound characteristics of a specific speaker are accumulated into the lifetime memory during the training phase, while the speech perceptron is trained to extract temporal sound characteristics and update the memory online when the speaker perceives speech. The learning memory is then used to interact with the mix input to add and filter the target frequency from the mix stream. Finally, the network is trained to minimize the reconstruction error of attendance speech. In this study, a single speaker's voice was extracted from a speech segment containing multiple speakers using an ASAM model, and then speaker recognition was performed using an LSTM neural network. In the LSTM network, MFCC, GFCC, and GBFB will be used to identify the three feature quantities and the results will be compared.**

1. はじめに

音声コミュニケーションは人間の生活に必要な一部であり、音声情報の表現内容は文字より直接的で豊かである。うるさい言語コミュニケーション環境は常に現実生活に現れている。私たちの脳は人間の耳に応じてさまざまな混合音を聞くことができ、気になる音を区別するのは簡単である。様々な音が混ざって、私たちの脳は気軽に興味のある音声を得ることができ、これは有名な「カクテル・パーティー」効果である。

カクテルパーティー問題は、1953年にイギリスの認知科学者 Cherry が選択注意 (Selective attention) のメカニズムを研究する時に提出した有名な問題である。この問題は他の話者や騒音に邪魔された場合、人類が目標の話者の言語を理解する過程の背後にある論理的基礎を探って、対象の話者の信号をフィルター処理できるスマートマシンをモデル化することである。簡単に言えば、カクテルパーティ問題は人間の複雑な聴覚環境における一種の聴覚選択能力である。この場合、人は興味がある音声に注意して他の背景音を無視できるが、聴覚モデルを計算すると騒音の影響が大きい。カクテルパーティー環境に柔軟に適応できる聴覚モデルの設計方法は、計算聴覚の分野で重要な問題である。音声認識、音声強調、話者認識、音声分離などの一連の重要なタスクにおける非常に重要な研究意義と応用価値がある。特に近年では、スマートデバイスとポータブルコンピューティングデバイスの爆発的な発展により、音声は、スマートコンピューティングデバイスとプラットフォームにアクセスするための最も重要な入り口の一つになっている。したがって、日常生活で最も典型的で一般的な複雑な聴覚シーンに直面して、カクテルパーティー問題に効果的に対処する方法は非常に重要である。言い換えると、カクテルパーティ問題のモデルの品質、つまり複雑な聴覚シーンのモデリング方法は、入力情報の完全性、キー情報が効果的にスクリーニングされるかどうか、干渉情報が無視されるかどうかに直接影響する。これは後続のタスクの成功に影響し、その重要性は当然自明である。前述のように、スマートデバイスの普及はカクテルパーティーの問題に前例のない課題と要件をもたらしましたが、同時に、人工知能の手法と分野の急速な発展はカクテルパーティー問題を解決するためのより良い機会ももたらした。

実際には、複雑な環境に直面した際の聴覚の選択と注意の能力は、人間の進化の過程で聴覚システムを形成するための驚くべき才能である。カクテルパーティー効果のメカニズムは非常に複雑であるが、人間が複数の音源を切り替えるのは非常に簡単であるため、このプロセスの存在さえ感じない。残念ながら、現時点では、インテリジェントマシンが人間と同じ理想的なパフォーマンスを達成することは困難である。しかし、半世紀以上の継続的な研究の後、カクテルパーティー問題の背後に隠された神経メカニズムはまだ明確ではないが、関連する研究は特定の結果を達成している。 たとえば、研究者は、人間の聴覚経路の形成と、神経

* Supervisor: Prof. Kakunobu Itou

伝達中の聴覚信号のエンコードについて明確に理解していた。一方で、今日の人工知能の方法とモデリング方式にとって、特にニューラルネットワークとディープラーニング手法では、人間の脳の過程における関連機構を参考にして、脳発見式のモデルを構築することが非常に有効な手段となっている。たとえば、畳み込みニューラルネットワーク（CNN）の設計プロセスは、人間の視覚の関連メカニズムへの参照であり、同様の計算モデルフレームワークを効果的に構築し、画像処理の分野で非常に顕著な進歩を達成した。したがって、聴覚に関するメカニズムを利用して、複雑な環境で話者認識を実行できる。

2. 関連研究

2.1. カクテルパーティーの問題

音声分離と聴覚選択は、聴覚システムを計算することによってカクテルパーティーの問題を解決するために、研究者は、計算による聴覚情景分析（CASA）、非負行列因数分解（NMF）からディープラーニングベースの方法まで、数十年間の多くの音声分離方法を提案した。辞書学習の最も代表的な例として、NMF はトレーニング中に各クリーニングソースを一連の話者関連辞書とアクティブ化に分解し、各ソースのアクティブ化を最適化してグローバルな最適化を実現する。しかしながら、これらの分解に基づく方法は以下の制限を有する：（１）これらの方法のほとんどは、反復法を通して大域的最適化を達成し、それはしばしば評価中に高い計算上の複雑さをもたらす。（２）背景や未知話者の雑音は、目的の辞書が事前に学習されていても分解モデルが高品質の分離を達成することを妨げる。（３）利用可能な辞書の総数が多いとき、たとえグループ希薄性ペナルティを導入しても、全ての可能なソースを再構築しようと試みることは実際的ではない。事前学習済み辞書の有用なサブセットの事前知識を選択することは分解に役立つが、複雑な聴覚シナリオでは、入力ソースは通常不確定で動的でさえある。最近のディープクラスタリング（ＤＣ）およびディープアトラクターネットワークはラベル置換の問題および出力寸法の不一致を経験しているが、どちらも混合入力内のすべての信号を分離しようとするために従来の計算フレームワークから逸脱することはなく、複雑な聴覚シナリオでは十分に機能しない可能性がある。人間の聴覚行動および選択的処理に関する最近の研究は、多人数話者環境では、聴衆の皮質活動は主にその人の話のスペクトルおよび時間的特性によって決定され、無人話者についてのみ弱いことを示している。関連性さらに、聴覚選択の神経生理学的研究は、人間の一次聴覚皮質が選択的周波数チャネルに同調するために周波数選択的注意フィルタを生成し、聴覚記憶における特定のオブジェクト表現がトップダウン注意の知覚を高めることを示した。正確さ上記の観察に基づいて、選択的注意および聴覚記憶を計算による聴覚モデルに統合することは、カクテルパーティー問題に対する実行可能な解決策となる。

2.2. 注意と記憶の強化モデル

最近のディープラーニングへの関心の高まりに伴い、多くの研究者は、テキスト、画像、音声を固定長ベクトルにマッピングするためにディープニューラルネットワークを使用することに焦点を当ててきた。これらの学習ベースの方法の主な利点は、それらが手作りの機能に依存しないことである。ただし、固定長ベクトル表現は、過去のオブジェクトを正確に記憶するには小さすぎることが多く、応答生成に重要な詳細を失う可能性がある。Jiaming Xu らは、注意と記憶を伴う単一の聴覚選択フレームワーク（ASAM と呼ばれる）を提案した。ASAM は最初にトレニニーグ段階の間に前の知識（すなわち、特定の話者の音響特性）を生涯の記憶に蓄積している間、音声受信機をトレニンーグして一時的な音響特徴を抽出して、そしてネットワークの記憶を更新する。次に、学習メモリを使用して混合物入力と対話し、混合物ストリームから目標周波数に参加し、それを除去する。最後に、ネットワークは、話している人の再構成エラーを最小限に抑えるようにトレニンーグされている。私はこの方法を使って実験した。

2.3. LSTM

Long Short Term Memory network（LSTM）は、長い依存関係の問題を解決するために設計された特別な RNN ネットワークである。 このネットワークは Hochreiter & Schmidhuber によって提案され、多くの人々によって改善して普及された。 LSTM はさまざまな問題を解決するために使用されており、今でも広く使用されている。

LSTM にはチェーン構造があるが、その繰り返し単位は、ネットワーク層が 1 つだけの標準 RNN ネットワークの単位とは異なり、内部に 4 つのネットワーク層がある。LSTM の中核はセルの状態である。 セルの状態はコンベアベルトのようなものであるが、セル全体に分岐がほとんどないため、RNN 全体に情報が流れるようになる。LSTM ネットワークは、ゲートと呼ばれる構造を介して、セルの状態の情報を削除または追加できる。 ゲートは、どの情報を渡すかを選択的に決定できる。 ゲートの構造は単純で、シグモイド層とドット積の組み合わせである。

LSTM の最初のステップは、セル状態から破棄する必要がある情報を決定することである。 操作のこの部分は、忘却ゲートと呼ばれるシグモイドユニットによって処理される。 h_{t-1} および x_t によって、0-1 の間のベクトルを出力する。このベクトルの 0-1 値は、セル状態 C_{t-1} のどの情報が保持または破棄するかを示す。0 は破棄することを意味し、1 は保持することを意味する。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

次のステップでは、セルの状態に追加する新しい情報を決定する。 このステップは 2 つのステップに分かれている：最初に、h_{t-1} と x_t を使用して、入力ゲートと呼ばれる操作で更新する情報を決定する。 次に、h_{t-1} および x_t を使用して、tanh レイヤーを介して新しい候補セル情報 $\tilde{C}_t$ を取得する。これらの情報は細胞の情報に更新されるかもしれない。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

古いセル情報 C_{t-1} は新しいセル情報 C_t に更新される。 更新のルールは、忘却ゲートの選択を通じて古いセル情報の一部を忘れ、入力ゲートの選択を通じて候補セル情報 $\tilde{C}_t$ の一部を追加して、新しいセル情報 C_t を取得する。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

セルの状態を更新した後、出力セルの状態特性を入力 h_{t-1} および x_t に従って判断する必要がある。ここでは、出力ゲートと呼ばれるシグモイド層を使用して判定条件を取得し、セル状態を tanh 層に通して-1 から 1 の間のベクトルを取得する。このベクトルは出力ゲートから得られた判定条件に乗算され、最終的にはこの RNN ユニットの出力が得られる。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

3. 実験目的

実際の環境は非常に複雑である。たとえば、話者認識システムが「カクテル・パーティー」に適用されると、多音源環境では、認識率が急激に低下し、多くの干渉源から目標音声を取得することは非常に困難である。騒音が含まれている環境から目標音声を識別するために、ASAM モデルを使ってカフェ、他の人の声、ホワイトノイズなどの騒音環境から単一音声を抽出する。この後、LSTM ネットワークを使って話者認識実験を行う。抽出された音声は雑音を含んでいる可能性があることを考慮して、話者認識部分では、騒音に抵抗できる GFCC 特徴量と GBFB 特徴量を LSTM ネットワークに入力して実験を行い、MFCC を対照として使用する。最後に、認識率を使用して、各特徴量の騒音の抵抗能力と認識効果を評価する。

4. 相関実験

LSTM を使用した話者認識実験では、MFCC、GFCC、GBFB を使用して実験を行う。 ここで主に行われた実験は GFCC および GBFB 特徴量抽出の実験である。

4.1. GFCC 特徴量の抽出

Gammatone フィルターバンクは、基底膜の周波数分割特性を十分にシミュレートでき、背景雑音をさらに抑制し、スピーカーの音声信号の明瞭度を改善できるが、音声信号の信号対雑音比を向上させることができない

入力音声信号は一連の Gammatone フィルターを通過し、音声信号は時間領域から周波数領域に変換される。 ここでは、16000 Hz のサンプリング周波数と 160 ms のフレームシフトを持つ 26 次の Gammatone フィルターが使用される。これは、短期音声特徴抽出に適用できる。 音声信号が上記のフィルターを通過すると、出力信号の応答 G_m(i) の式は次のようになる。

$$G_m(i) = [|g|(i,m)]^{1/2}, i = 0, 1, ..., N-1; m = 0, 1, ..., M-1 \tag{7}$$

ここで、N = 26 はフィルタのチャネル数であり、M はサンプリング後のフレーム数である。このように $G_m(i)$ は、入力信号の周波数領域での分布変化を表す行列を構成する。

4.2. GBFB 特徴量の抽出

GBFB は、Gabor フィルターを使用して時間および周波数ドメインにわたって音響特徴を抽出し、高次元 GBFB 特徴を時間および周波数ドメインの異なる部分空間にマッピングする方法である。これにより、雑音成分を除去し、堅牢な特徴を保持する。 GBFB 機能は、いくつかの一般的な音響特徴よりも雑音の多い環境での堅牢性が優れている。Gabor 特徴抽出は窓付き Fourier 変換に基づいて実現され、Gabor 変換によって時間領域をまたいで特徴情報を抽出することができるので、より有用な特徴情報を得ることができる。

音響特徴を抽出するための二次元 Gabor 関数の定義は以下の通りである:

$$g_{u,v}(n,k) = \frac{\gamma^2}{\sigma^2} \cdot e^{-\frac{\gamma^2 \varepsilon^2}{2\sigma^2}} \cdot (e^{i\gamma\varepsilon} - e^{-\frac{\sigma^2}{2}}) \tag{8}$$

ここで、$\gamma = k_v e^{i\phi}$ はフィルタの方位とスケールを決定し、$\varepsilon(n,k)$ は FFT によって得られた音響スペクトルサンプルポイントを表し、$\phi = u(\pi/k)$ 、$k_v = 2^{-((v+2)/2)} \cdot \pi$ は、u を変化させることにより、v は Gabor フィルタセットの方位とスケールを調整することができる。その後、処理を簡略化し、得られた特徴量をサンプリング周波数 16000Hz の 26 次メルフィルターに入力し、各フレームの MFCC パラメーターを計算して特徴量を求めた。

5. 雑音抑制実験

5.1. データセット

このモデルはより少ないデータでトレーニングできるため、ラベルも調整された。トレーニングセットは 10 人の音声があって。一人に 7 つの音声ファイルがある。テストセットは一人に 5 つのセグメントしかない。各人の各セグメントは他の 9 人のセグメントの中からランダムに一つを選んで雑音として、最後にラベルを付ける。パラメータを調整するためのデータセットは一人に 3 つの音声セグメントがある。各音声セグメントは他の 9 人の各音声セグメントでパラメータを調整し、各音声セグメントは 27 回調整することができる。未知の話者のデータセットには 5 人の音声セグメントがあり、テストセットと既知セットに分けられる。テストセットの各音声セグメントは、他の 4 人のセグメントでそれぞれテストされ、自分の 10 個の音声セグメントを刺激として使用する。このようにして、各セグメントを複数回パラメータを調整またはテストでき、データセットが大幅に削減された。

5.2. 実験内容

実験は Libri データセットを用いて行った。2 人の話者の音声を線形に混ぜる。混合音は、認識目標によって、トレーニング、パラメータ調整、およびテストセットに分けられ、不明なスピーカーはそれぞれ刺激として 10 のクリーンな音声を保持する。すべてのデータは 8000Hz にリサンプリングされている。振幅スペクトルは入力特徴として用いられる。３２ｍｓのウィンドウ長、１６ｍｓのジャンプサイズおよび結果ウィンドウを有する短時間フーリエ変換（ＳＴＦＴ）計算が使用される。それぞれが 100 個の小さなバッチを含む、小さなバッチに対して 8 つのサンプルをランダムに生成する。少なくとも 10 期間で、損失を検証するためのトレーニングは少なくとも 150 回行われた。このアーキテクチャでは、2 層の BiLSTM が使用され、ハイブリッドエンコーダ用に各方向に 300 の隠れユニットがあり、音声受信機およびメモリ用に各方向に 20 の隠れユニットがある 2 層の BiLSTM がある。Ｔ-Ｆ埋め込みベクトルおよび格納ベクトルの次元はすべてｄ＝４０に固定されている。聴覚選択の結果は GNSDR を用いて定量的に評価された。

表 1. テスト結果

	GSDR	GSIR	GSAR	GNSDR
valid spk 2	7.27	10.75	9.53	7.28
test spk 2	7.84	11.44	10.07	7.84
test spk 3	1.51	3.55	7.43	4.89
test spk 2 bg noise	-1.06	1.27	5.98	4.81
unk spk supp_time=0	1.43	2.75	5.32	1.43
unk spk supp_time=4	3.06	4.96	6.81	3.06
unk spk supp_time=16	2.41	4.25	6.22	2.41

テストの結果はあまりよくなかった。予想との差が大きいである。メモリが限られているため、batch size が小さすぎるように調整された結果であると考えた。batch size を少し大きくして、もう一度実験した。batch size は 16 に大きくなり、評価用の batch size は 48 に大きくなった。結果は以下の通りである。

表 2. テスト結果

	GSDR	GSIR	GSAR	GNSDR
valid spk 2	8.02	11.93	9.82	8.02
test spk 2	8.67	12.72	10.37	8.67
test spk 3	2.12	4.52	6.95	5.50
test spk 2 bg noise	-1.13	0.86	6.29	4.74
unk spk supp_time=0	1.69	2.73	5.18	1.69
unk spk supp_time=4	3.56	5.49	7.21	3.56
unk spk supp_time=16	3.18	4.82	6.67	3.13

ハイパーパラメーターを数回調整した後、より良いモデルが得られた。

5.3. ASAM モデルによる雑音除去

二人の声を同じデシベルで合わせて、wav ファイルを作って、モデルをテストする。wav ファイルはモデルを通して一人の声を除いた wav ファイルを出力する。

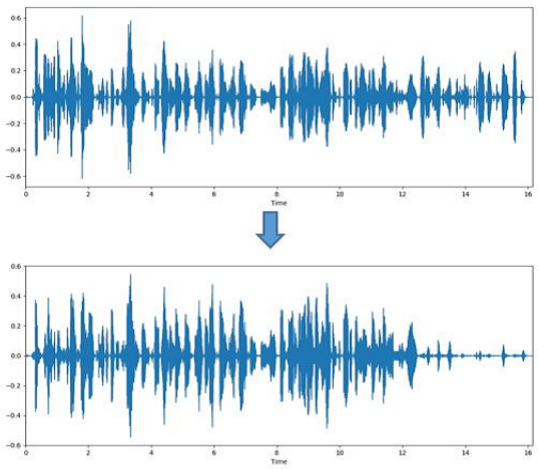

図 1. 処理前後の振幅スペクトル

モデルの処理を通して、その中の一人の声が弱くなり、もう一人の声が少し強くなった。弱くなった音声はぼやけても聞こえるが、モデル効果は限られている。

6. 音声認識実験

matlab で LSTM ネットワークを作って話者認識実験を行った。データセットには 4 人の声、男性 2 人、女性 2 人が含まれている。トレーニングセットには一人の声が 100 の音声ファイルがあって、テストセットには一人の声が 20 の音声ファイルがある。各ファイルの時間は約 12 秒である MFCC，GFCC，GBFB の三つの特徴量を用いてカフェ，人間の声、ホワイトノイズをそれぞれ実験した。

6.1. 雑音がない場合

表 3. 雑音がない認識率

	MFCC	GFCC	GBFB
認識率	0.9286	0.6786	0.8571

騒音がない場合、MFCC の識別率が最も高いため、GBFB の特徴は純粋な音声音響特性を抽出する際、絶対値の小さい成分を騒音処理と間違えてしまうため、きれいな音声に一定の程度の損傷があり、GBFB の純粋な音声環境における識別率は他のいくつかの特徴よりも著しく向上しておらず、場合によっては他の特徴よりも低い場合もある。

6.2. 雑音がある場合

表 4. カフェ環境での認識率

	MFCC	GFCC	GBFB
未処理	0.25	0.50	0.39
雑音除去後	0.33	0.47	0.46

表 5. 二人の声での認識率

	MFCC	GFCC	GBFB
未処理	0.36	0.43	0.29
雑音除去後	0.45	0.49	0.28

表 6. ホワイトノイズでの認識率

	MFCC	GFCC	GBFB
未処理	0.21	0.43	0.29
雑音除去後	0.54	0.29	0.31

騒音がある環境では，3 つの特徴量の識別率が大幅に減少し、GFCC も一定の雑音耐性を持っていて、下げ幅が最小である。ASAM モデルの処理によって識別率は 6%-10%上昇した。しかし、一部の識別率が逆に下がっていることもある。モデルは音を抽出することができたが、騒音と人の声を区別することができず、背景騒音が増幅され、モデルはその中のいかなる音も完全に除去できず、最終的にはいくつかの特徴の識別率が低下した。

7. まとめ

ASAM モデルにはノイズの抑制に一定の効果があるが、対応する人の声がトレーニングセットに保持されることを保証できず、認識率が低下する可能性がある。 さらに、モデルのトレーニング時に他のノイズが追加されないため、モデルは雑音をうまく処理しない。 話者認識部分では、MFCC は雑音がない場合に認識率が最も高く、雑音を追加した後、GFCC は一定の雑音耐性を持っているパフォーマンスを示す。 その後、他の雑音がモデルトレーニングに追加され、雑音の処理能力が向上させる。

参考文献

[1] HINTON G,DENG L,YU D,et al.Deep neural networks for a coustic modeling in speech recognition:The shared views of four research group [J].IEEE Signal Processing Magazine,2012,29(6):82-97.

[2] YIN Rui-Gang, WEI Shuai, LI Han, YU Hong.Introduction of Unsupervised Learning Methods in Deep Learning

[3] ZHAO Yan, Lu Liang , ZHAO Li.Research on Speaker Identification Based on Improved Deep Neural Network

[4] Jiaming Xu, Jing Shi, Guangcan Liu, Xiuyi Chen, Bo Xu Modeling Attention and Memory for Auditory Selection in a Cocktail Party Environment

[5] LIN Haibo , WANG Kejia New algorithm for auditory feature extraction

[6] GOU Xinke , XU Gaopeng Research on Speech Recognition Robustness based on Gabor Filter

[7] Sepp Hochreiter , jurgen Schmidhuber LONG SHORT-TERM MEMORY

[8] Dominic Masters , Carlo Luschi REVISITING SMALL BATCH TRAINING FOR DEEP NEURAL NETWORKS

[9] Zhou Chen , Jinyu Li , Xiong Xiao CRACKING THE COCKTAIL PARTY PROBLEM BY MULTI-BEAM DEEP ATTRACTOR NETWORK

[10] Ziang Xie , Sida I.Wang , Jiwei Li ATA NOISING AS SMOOTHING IN NEURAL NETWORK LANGUAGE MODELS

[11] Zekeriya Uykan On the “SIR”s (“Signal”-to-“Interference ”-Ratio) in Discrete-Time Autonomous Linear Networks

[12] Cao Jingjing , Xu Jieping Research on Multi-noise-robust Auto Speech Recognition

[13] Yang Dali , Xu Mingxing Speech Recognition Study in Noisy Environment