Deep Clustering：話者・言語非依存なシングルチャネル音声分離技術

相原　龍*
Ryo Aihara
ウィシャーン ゴードン†
Gordon Wichern
ルルー ジョナトン†
Jonathan Le Roux

Deep Clustering：Speaker and Language Independent Single Channel Speech Separation Technology

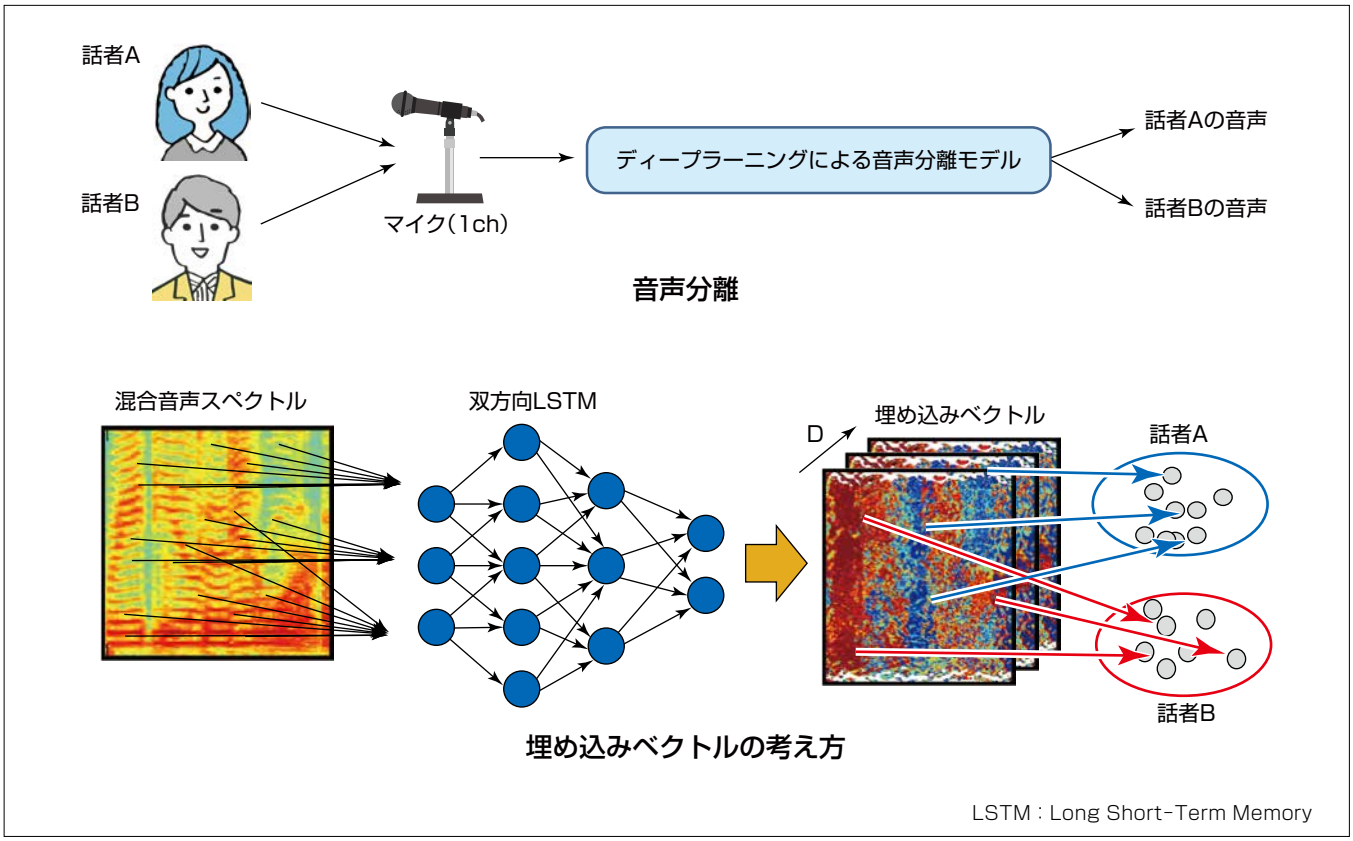
要　旨

音声分離とは，複数話者の重畳音声からそれぞれの話者の音声を抽出する技術である。

ヒトは，二つの耳への音声の到来時間差から音源の位置を推定し，さらには視覚情報などを統合的に用いることで，複数話者の重畳音声から目的話者の音声を聞き取っていると考えられる。機械にとって，音声同士を分ける音声分離は，音声とノイズとを分離する音声強調と比較して難しい課題であり，特に音源の位置情報を得ることができないマイク１本での話者・言語非依存なシングルチャネル音声分離は困難であるとされてきた。

Deep Clustering(DC)は，ディープラーニング(深層学習)とクラスタリングを組み合わせることで，世界で初めて(注1)話者・言語非依存でありながら高精度なシングルチャネル音声分離を実現した。英語しか学習していない音声分離モデルで，未学習の日本語やそのほかの言語も事前登録なしに分離が可能であり，２話者の同時発話であれば90％程度，３話者では80％程度の再現率が得られる。公共空間やクルマなど，複数のヒトが同時多発的に発話する環境で音声認識を行う際には，背景ノイズだけでなくユーザー以外の音声によって認識率が大幅に低下することがあり，音声分離技術によって多様な環境での音声認識精度向上が期待できる。

(注１)　2017年５月24日現在，三菱電機調べ



音声分離と埋め込みベクトルの考え方

DCは，マイク１本だけで話者・言語に依存することなく音声分離ができる。ディープラーニングによる音声分離モデルでは，英語話者の音声だけを学習しており，事前に分離する話者の声を登録することなく音声分離が可能である。この技術の要になっているのは，埋め込みベクトルの考え方である。混合音声スペクトルからD次元の埋め込みベクトルを推定することで，古典的なクラスタリング手法で分離が可能になる。

*三菱電機㈱ 情報技術総合研究所(博士(工学))
†Mitsubishi Electric Research Laboratories(Ph.D.)

1. ま え が き

大勢の人がそれぞれに雑談している環境でも，ヒトは自分の興味のある話題や他人との会話を自然に聞き取ることができる。このような音声の選択的聴取を“カクテルパーティー効果”と呼び，イギリスの心理学者Cherryによって提唱された[1]。Cherryは，被験者の左右の耳に同時に異なる音声を流し，指定された方の音声をシャドーイングする実験を行った。機械やロボットにカクテルパーティー効果を持たせようという試みは，20年以上にわたって多くの音声信号処理研究者によって研究されてきた。

一般に，ヒトは二つの耳を持ち，両耳への音の到来時間差を利用して音源の方向を推定していると考えられる。複数のマイクを用いたマルチチャネル信号処理による音源分離は既に多くの場面で実用化されており，目的音源の方向に指向性を向けるビームフォーミングがその代表的な手法である[2]。しかしながら，これらの手法は音源同士が近くに存在している場合には分離が困難であり，当然ながらマイクが一つしかないシングルチャネル環境では動作しない。

ディープラーニングの登場は機械学習の分野に飛躍的な進歩をもたらしたが，音声信号処理でもそれは例外ではない。ノイズ環境下音声からノイズを除去する“音声強調”では，ディープラーニング技術によってシングルチャネル音声強調が可能になった[3]。これはディープラーニングが，音声強調に適した特徴量の抽出と，音源情報のモデリングを一体で行うことができるためである。しかしながら，ノイズではなく，複数話者の重畳音声を話者ごとに分離する“音声分離”は，ディープラーニングをもってしても困難な課題であった。これは，ノイズと音声はその特徴が異なるのに対して，ヒトの音声同士はその特徴が類似しているためである。

DCは，ディープラーニングとクラスタリングを応用することで世界で初めて話者・言語非依存でありながら高精度なシングルチャネル音声分離を実現した[4]。英語しか学習していない音声分離モデルで，日本語やそのほかの言語も分離が可能であり，２話者の同時発話であれば90％程度，３話者では80％程度の再現率が得られている。この技術によって複数人が同時に話す環境でマイク１本での音声分離が可能になったため，多様なシーンでの音声認識の精度向上が期待されている。

本稿では，音声強調と比較してシングルチャネル音声分離がなぜ難しいのかを述べた後，DCの基本的なアルゴリズムとその発展的手法や低遅延化手法について述べる。

2. 音声分離の難しさ

本稿では，音声以外のノイズが重畳した音声からノイズを除去する技術を“音声強調”，音声同士が重畳した音声からそれぞれの話者の音声を分離する技術を“音声分離”として区別する。音声強調の場合は，抽出したい音声と除去したいノイズの特徴が異なることから，ノイズ重畳音声，正解となるクリーン音声及び除去したいノイズを与えてモデルを学習すればよい。しかしながら，音声分離の場合は特徴が似ているヒトの音声同士を分離するため，音声強調と同じような直接的分離モデルの構築が困難であった。

その原因の一つに，Permutation問題がある[4][5]。**図１**に，話者非依存な２話者音声分離モデルを学習する例を示す。話者非依存なモデルを学習する場合，様々な話者ペアの音声を重畳し，この重畳音声と，正解となるそれぞれの話者の非重畳音声とを，学習データとして分離モデルに入力する必要がある。まず，**図１**の左側のように話者(A，B)の混合音声のサンプルを学習データとして入力し，左側に話者A，右側に話者Bを出力するように正解データを与える。次に話者(A，C)の混合音声サンプルを学習する場合，先ほど話者Aの正解データを左側に配置したため，**図１**中央のように話者Cの正解データは右側に配置にすることが自然であろう。しかし，話者(B，C)の混合音声サンプルを学習する場合，これまで両方とも右側に配置してきた話者の正解データを左側に配置すると，それ以前の学習との一貫性が失われてしまう。このように，２話者混合音声を３名の組合せで学習する場合でも正解データの配置方法が６通りも考えられ，混合音声から正解の分離音声を直接学習することが困難であることが分かる。

3. DCによる音声分離

DCは話者混合音声から正解音声を直接推定するのではなく，ディープラーニングを用いて“音声を分離しやすい特徴量”を推定することでシングルチャネル音声分離を可能にした。このような特徴量を“埋め込みベクトル”と呼ぶ。

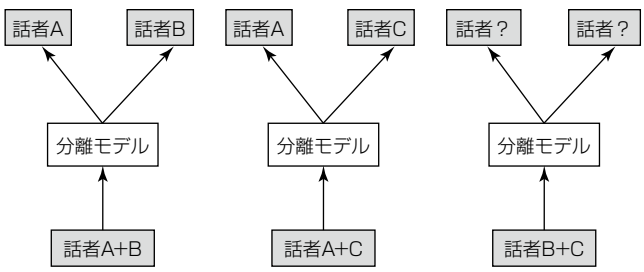


図１．Permutation問題

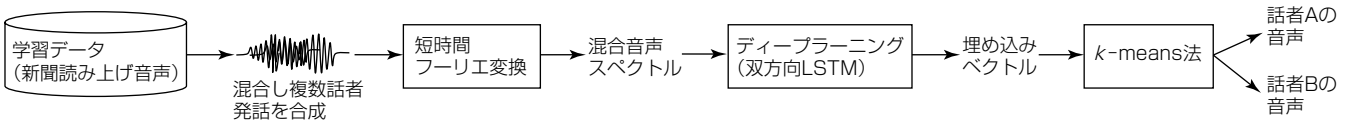


図２．DCの処理フロー

図２にDCの概要を示す。

音声は空気の振動であり，一次元の波形データとして表現される。波形データを短時間フーリエ変換することで，時間と音の高さを表す周波数の二次元データとして表現できる。このような時間-周波数表現をスペクトルと呼ぶ。２話者それぞれの音声スペクトルと，２話者音声を重畳させた音声スペクトルを比較すると，混合音声スペクトルのある時間tとある周波数fの要素点(t, f)で，どちらの話者が支配的か（どちらの話者のパワーが大きいか）を計算できる。混合音声スペクトルの全要素数をI（時間長をT，周波数長をFとするとき，$I = T \times F$）とすると，音声分離は，混合音声スペクトルの全I個の要素点について，どの混合話者に属するかを決定する問題とみなせる。

しかしながら，混合音声スペクトルの全要素について，どの話者に属するかを直接推定することは困難である。**２章**で述べたように，Permutation問題によって正解ラベルの与え方が一意に決定できないためである。そこで，全ての混合音声スペクトルの要素点それぞれに対して，ディープラーニングを用いて数次元の埋め込みベクトルを推定する。ある混合音声スペクトルの要素点同士が同一話者に属する場合，対応する埋め込みベクトル同士の距離は近くなるよう，逆に異なる話者に属する場合は距離が遠くなるような基準で埋め込みベクトルを推定する。“ある混合音声スペクトルの要素iがどちらの話者に属するか”は正解の与え方がPermutation問題によって一意に定まらないが，“ある混合音声スペクトルの要素同士が同じ話者に属するか否か”は，０か１かで表現された$I \times I$の二次元の正解ラベルデータで表すことができるため，深層モデルでの学習が可能である。

このようにして推定したI個の埋め込みベクトルは，混合音声に含まれる話者に対応するクラスタを形成していると考えられる。k-means法のような簡単な教師なしクラスタリング手法を用いて，埋め込みベクトルを話者数分のクラスにクラスタリングできる。クラスタリングされた埋め込みベクトルに基づいて混合音声スペクトルを抽出し，逆フーリエ変換で音声波形に戻せば，単一話者だけの音声信号を得ることができる。

ここで注意しなければならないのは，埋め込みベクトルは，“注目する混合発話内の，対応するスペクトルの要素同士が同じ話者に属するか否かに基づいている”ため，話者固有の声質を表現しているわけではないことである。埋め込みベクトルは，あくまで“混合音声から音声を分離しやすい特徴量”であるため，話者（A，B）の重畳音声から推定された話者Aの埋め込みベクトルと，話者（A，C）の重畳音声から推定された話者Aの埋め込みベクトルは同じである保証はなく，重畳されている音声によって，埋め込みベクトルは分離しやすいように変化すると考えられる。

埋め込みベクトルの推定には再帰型ニューラルネットワークの一種である双方向LSTMを用いることで，混合音声スペクトルの要素同士の前後関係を考慮した推定が行われる。双方向LSTMは発話全体を録音する必要があるオフライン型の処理であり，この低遅延化手法については**５章**で述べる。

4．より直接的な音声分離手法

DCは埋め込みベクトルを介してクラスタリングを利用することでシングルチャネル音声分離を実現したが，より直接的な音声分離手法も存在する[(4)(6)]。**図３**にその概要を示す。２話者音声分離モデルを学習するため，話者（A，B）の混合音声のサンプルを学習データとして入力する場合を考える。音声分離モデルの二つの出力に対して，正解データの配置は（A，B）と（B，A）のような二つのパターンが考えられる。両方のパターンとの誤差をそれぞれ計算し，誤差が小さい方を正解としてモデルに伝播（でんぱ）させ，モデル学習する。これを全学習データに対して繰り返し行うことで話者非依存な音声分離モデルを学習できる。これは，深層モデルが出力話者の順列（＝permutation）を全て計算し，どの順番で出力するかを，入力された発話データごとに判断・決定していることに相当する。

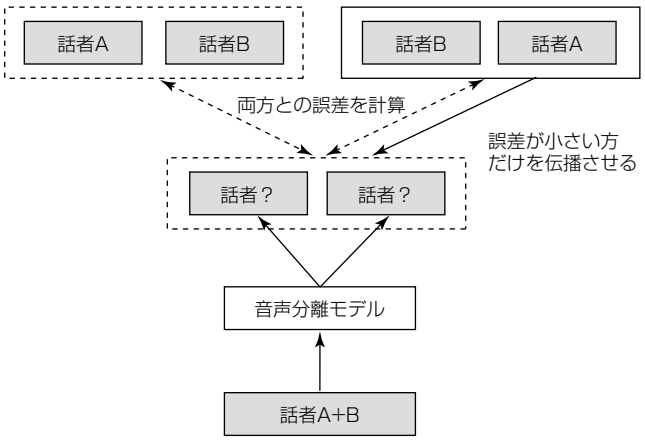


図３．より直接的な音声分離モデル

この直接的な音声分離手法とDCを組み合わせることで，更に高精度な音声分離モデルを構築できる[7]。この手法では，先に述べた直接的な音声分離と同時に，DCの埋め込みベクトルの推定を行う。モデルの学習時には，**3章**で述べたDCの誤差関数と直接的な音声分離モデルの誤差関数とを，重みを付けた上で足し合わせる。推定された埋め込みベクトルはモデルの学習だけにしか用いないが，埋め込みベクトルの推定によってモデルの汎化性能が向上し，より高精度な音声分離モデルが得られる。

5．低遅延化手法

ここまで述べた手法では，音声分離モデルに双方向LSTMを利用している。双方向LSTMの内部では，時間軸方向にforward層とbackward層とがあり，前者は時系列の順方向の依存関係を，後者は逆方向の依存関係をとらえる。この構造のため，発話全体を入力するオフライン処理となり，音声分離の開始までに発話長以上の遅延が発生する。双方向性を取り除き，過去との依存関係だけをみる通常のLSTMでは，分離精度が大幅に劣化することが実験的に明らかになっている[8]。

双方向LSTMのブロック処理によって，発話長以上の遅延が発生していた処理開始遅延を0.6秒にまで削減することに成功した[8]。**図4**にその概略を示す。単純に発話をブロックに区切るだけでは，ブロック間の依存関係を考慮できない。そこで，各ブロックに一定時間長の補助ブロックを持たせる。この補助ブロックは時間方向の次のブロックと重なりを持たせており，forward層に関してはメインブロックからの情報を伝播させ，backward層に関しては補助ブロックから伝播させた依存関係を考慮する。このような処理を行うことで，forward層に関しては双方向LSTMとほぼ同等，backward層に関しては最低でも補助ブロックの長さ分の情報を見ることができる。メインブロックと補助ブロックのブロック長によって分離精度と処理開始遅延が変化する。

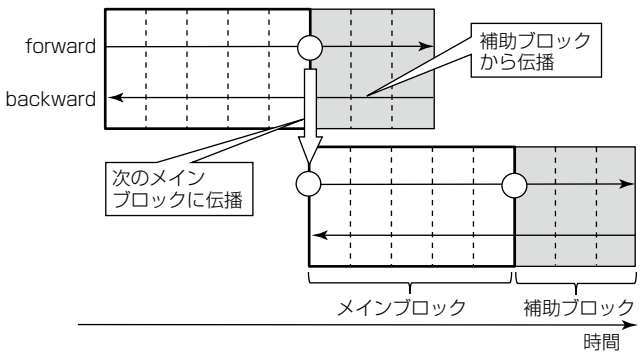


図4．低遅延化手法

さらに，従来の双方向LSTMの音声分離モデルを用いて，低遅延音声分離モデルの分離精度を向上させることができる。このように，あるモデルの情報を他のモデルに伝播させる技術を“知識蒸留”と呼び，通常の教師あり学習と組み合わせることで，低遅延モデルでも，従来の双方向LSTMとほぼ同等の，２話者発話の分離精度90％程度を実現した。

6．む　す　び

世界初のシングルチャネル音声分離技術であるDCについて，基本的なアルゴリズムとその拡張，さらには低遅延化手法について述べた。次に，ここまでで触れることができなかったDCの発展的手法について述べる。

ビームフォーミングを含む多チャネルのマイクアレーによる音声信号処理は，既に多くのシーンで実用化されている。DCの多チャネル拡張は参考文献[9]で提案しており，複数の２チャネルDCで求めた埋め込みベクトルを結合することで，多チャネル音声分離の精度向上が得られた。この手法が実用化されれば，更に安定した複数話者発話の同時発話分離が期待できる。音声から雑音を除去する音声強調と，音声同士を分離する音声分離を同時に行う手法も検証しており[10]，今後は更に多様な環境下での音声分離が実現する可能性がある。

以上のように，ディープラーニングの導入によって音声信号処理技術が利用できる可能性が，これまでの限られた環境からノイズ環境下，複数話者環境下へと大幅に広がりつつあり，現在これらの技術の実用化に向けて開発を進めている。

参 考 文 献

(1) Cherry, C.：On Human Communication, MIT Press (1966)
(2) Farrell, K., et al.：Beamforming microphone arrays for speech enhancement, Proc. IEEE ICASSP, 285～288 (1992)
(3) Wang, Y., et al.：Towards scaling up classification-based speech separation, IEEE Trans. on Audio, Speech, and Language Processing, **21**, No.7, 1381～1390 (2013)
(4) Hershey, J. R., et al.：Deep clustering：Discriminative embeddings for segmentation and separation, Proc. IEEE ICASSP, 31～35 (2016)
(5) Chen, Z.,：Single channel auditory source separation with neural network, Ph. D. dissertation, Columbia Univ. (2017)
(6) Yu, D., et al.：Recognizing multi-talker speech with permutation invariant training, arXiv preprint (2017)
(7) Wang, Z.-Q., et al.：Alternative objective functions for deep clustering, Proc. IEEE ICASSP, 686～690 (2018)
(8) Aihara, R., et al.：Teacher-student deep clustering for low-delay single channel speech separation, Proc. IEEE ICASSP, 690～694 (2019)
(9) Wang, Z.-Q., et al.：Multi-channel deep clustering：Discriminative spectral and spatial embeddings for speaker-independent speech separation, Proc. IEEE ICASSP, 1～5 (2018)
(10) Wichern, G., et al.：WHAM!：Extending speech separation to noisy environments, Proc. ISCA Interspeech, 1368～1372 (2019)