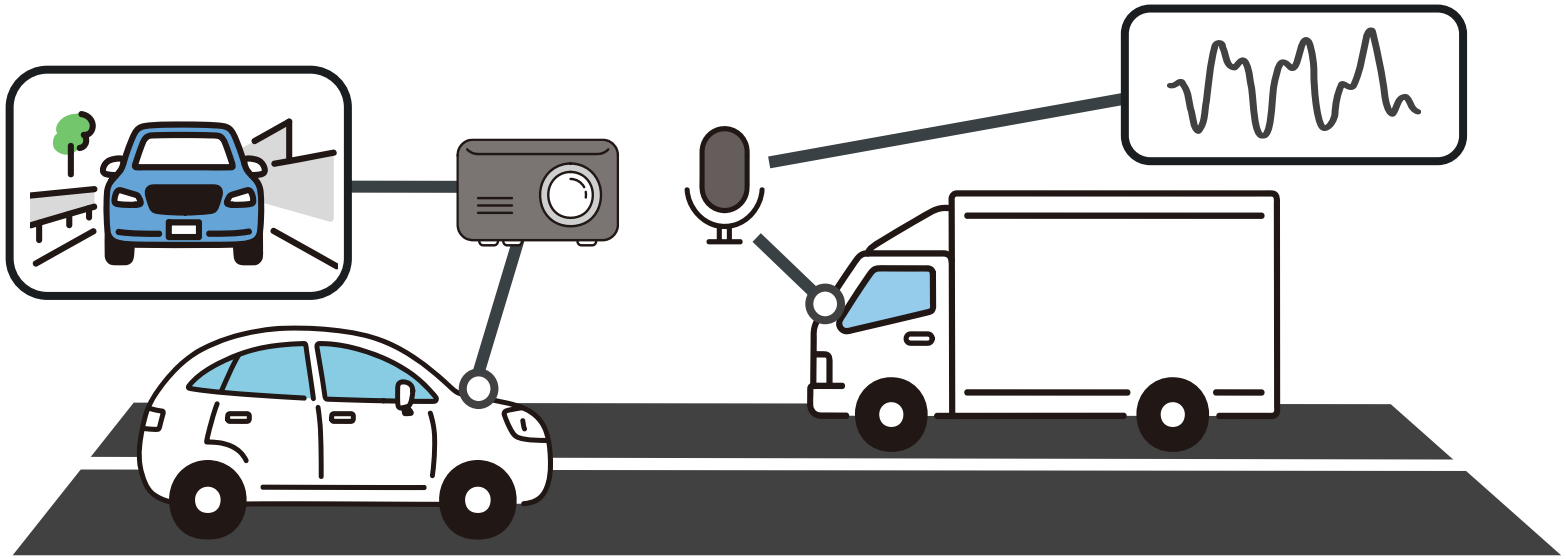
深層学習を用いた道路交通モニタリングシステムの開発

医療福祉機器開発工学コース 鄭研究室
二年 近藤 空哉

研究背景

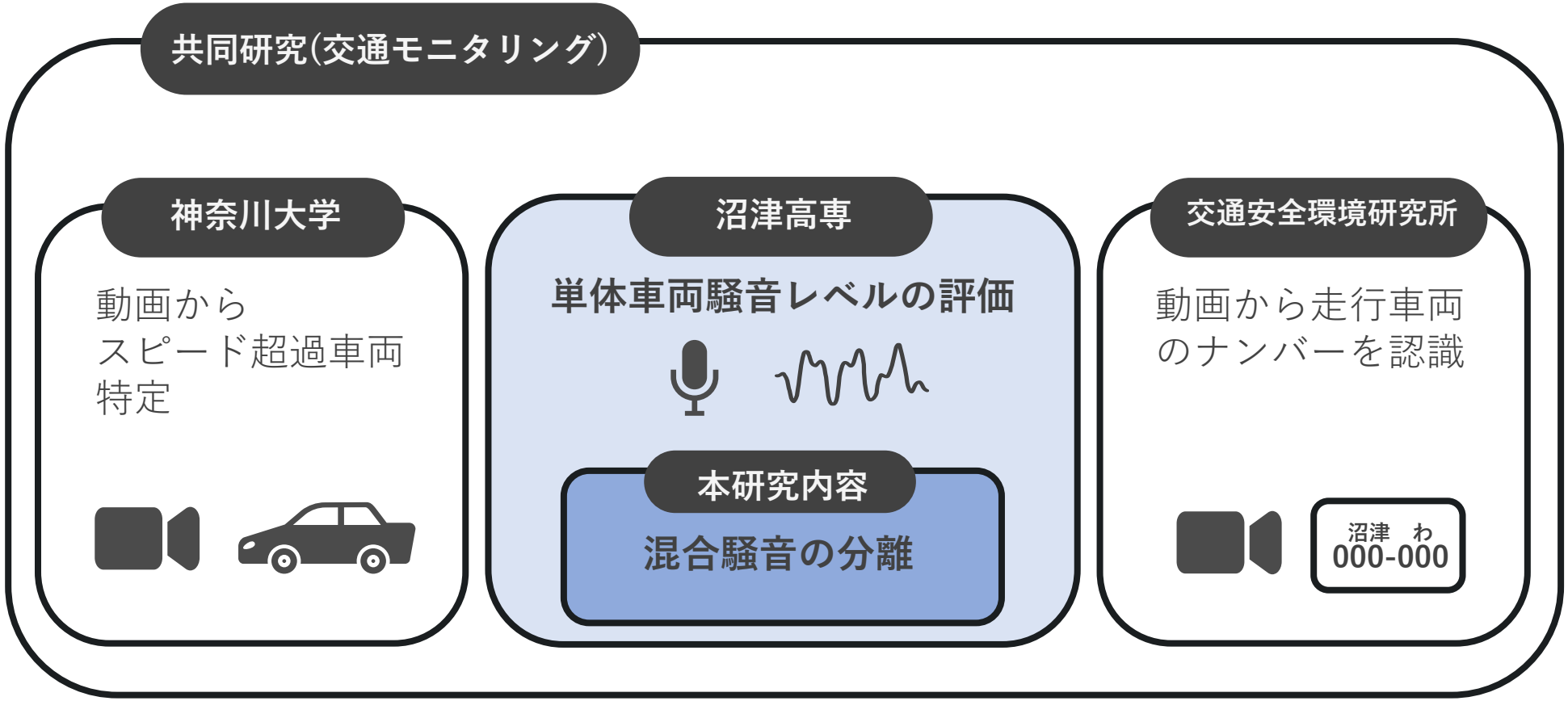
研究背景

来たる自動運転時代や、交通と暮らしへの注目から
交通インフラの再整備の重要性が見直されている



道路交通データを取得し、実態を正しく把握することで、交通課題
解決や自動運転車両による交通網システムの制御につながる
車両情報を抽出する道路交通情報取得システムの確立が求められる

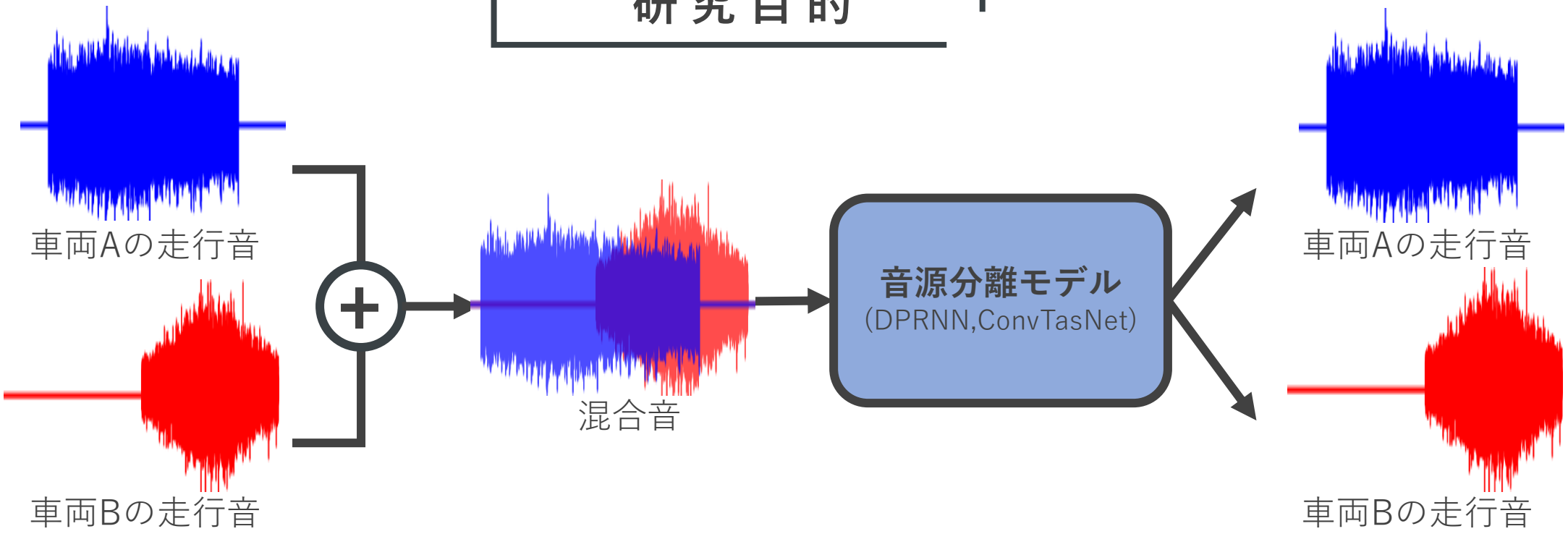
本研究と共同研究の関係



神奈川大学・交通研究所と共同でモニタリングシステムを構築
本研究は騒音車両特定のための**複数車両走行音を分離するモデル開発**を目標としている

研究目的

研 究 目 的



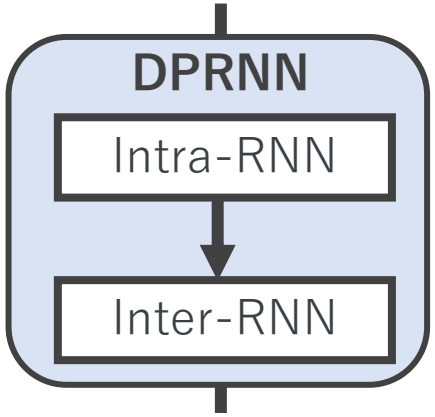
▲ 車両A・Bの混合走行音を単一走行音に分離するモデル

Dual Path RNN(DPRNN)という**話者分離等に活用されるモデル**を転用
Conv-TasNet(従来)を超える精度の**音源分離モデルの開発**

モデルの選出理由

Conv-TasNet

・従来手法の音源分離モデル
・TCNを用いて分離をする
・長期記憶性がやや弱い



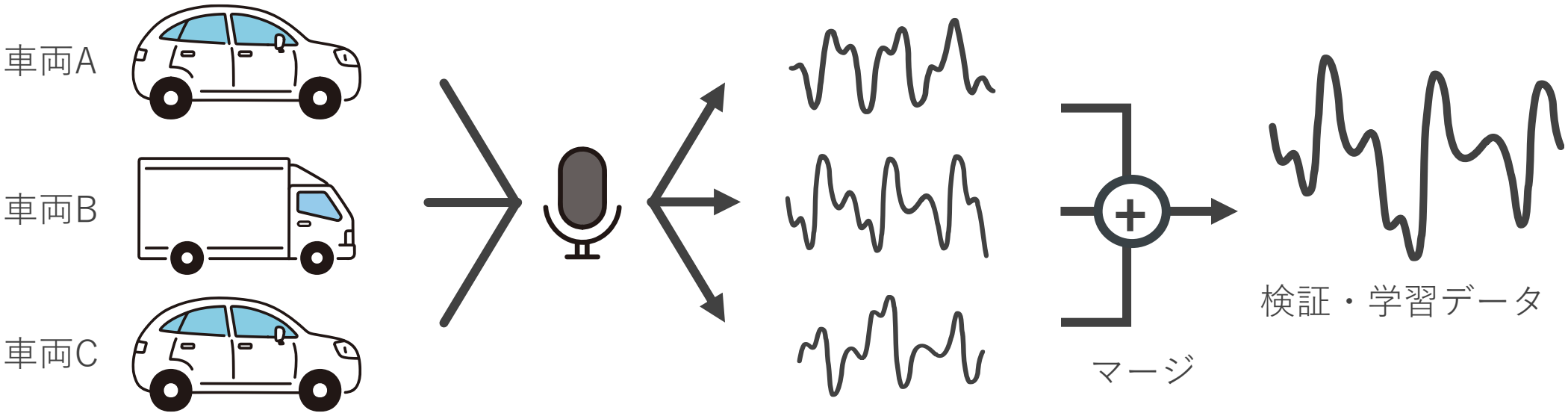
Dual Path RNN

・入力の圧縮率が高い
・長期記憶性が非常に強い

実験方法

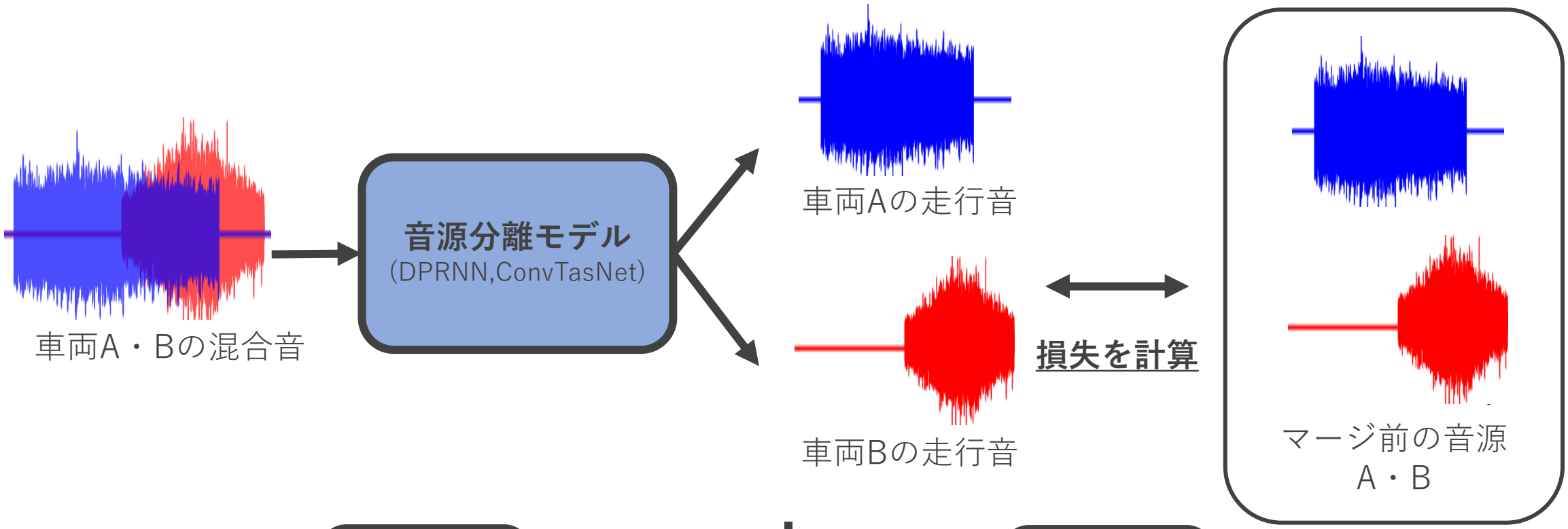
使用するデータ

学習・検証用のデータは録音した車両走行音を使用
データは計算負荷の軽減のため22000Hzにリサンプリングする



道路沿いにマイクを設置し、車両走行音を録音
単体車両走行音のみを切り抜き、それぞれをマージし
混合走行音を作成したものを利用する

実験方法



実験方法

Conv-TasNetとDPRNNに混合音を入力、それぞれの出力の**損失を比較**する

評価方法

マージ前の単体の音源と出力の分離音源の差をSI-SNRで計算したものを評価基準とする

損失関数

$$\mathbf{SI-SNR\_SP} = -\mathbf{SI-SNR} + \mathbf{alpha} * \mathbf{volume\_loss}$$

本研究で使用する損失関数　　従来の話者分離で使用されていた損失関数

$$s_{target} = \frac{<\hat{s},s>s}{\|s\|^2},\ e_{noise} = \hat{s} - s_{target}$$

$$SI-SNR = 10\log_{10}|\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}|$$

$$volume\_loss = \log_{10}|\frac{||S||^2 - ||\hat{s}||^2}{2e^{-5}}|$$

※$\hat{s}$：分離後の音源　s：分離前の音源

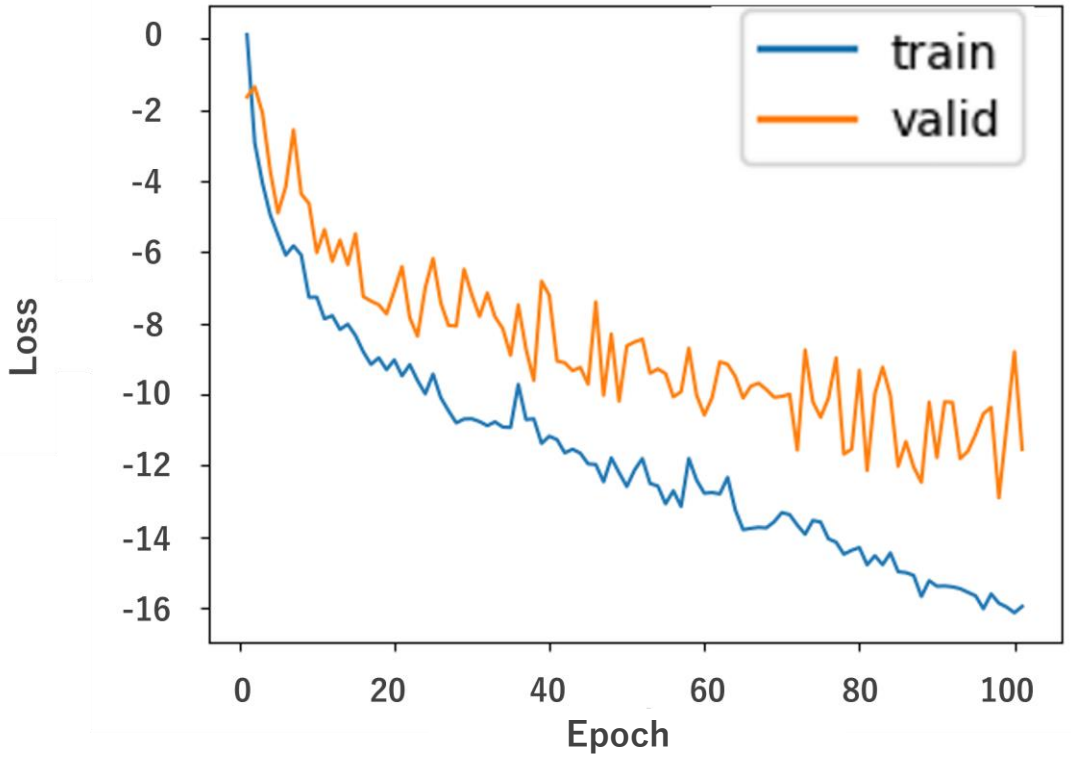
従来の損失関数SI-SNRでは音圧の損失を学習することができないため、

音圧の損失を考慮した損失関数SI-SNR_SPを提案・使用する

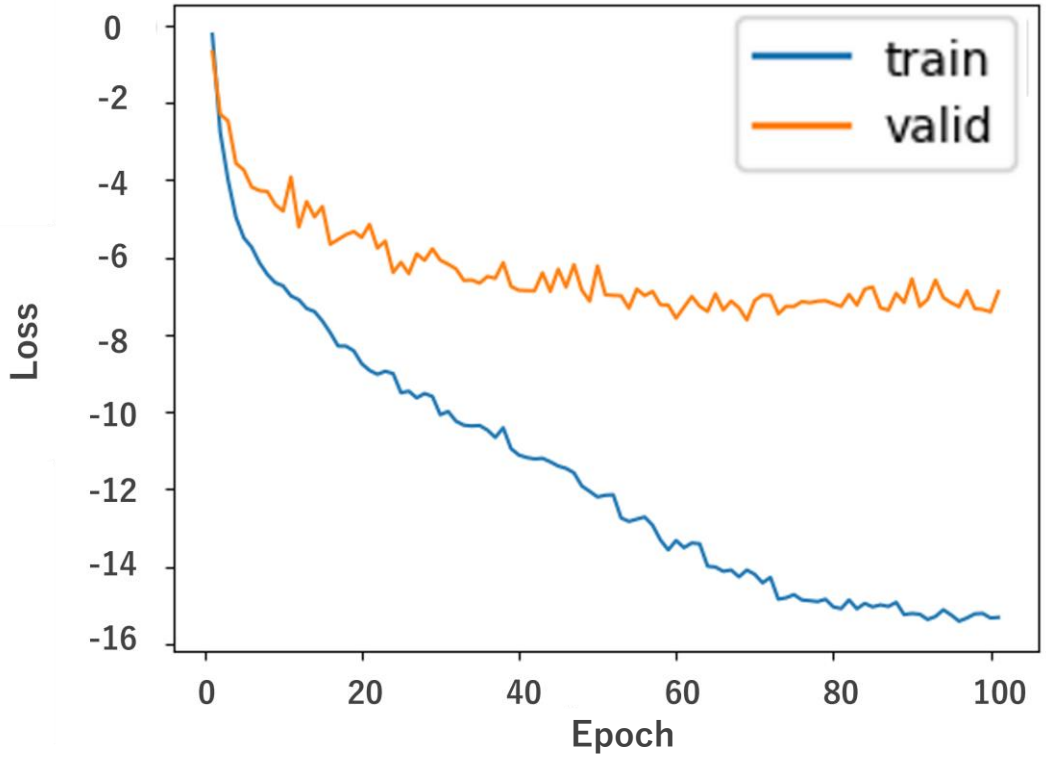
学習結果

学習の損失

DPRNNとConv-TasNetの学習曲線(n=2)について考える



DPRNN 2-mixed



Conv-TasNet 2-mixed

両モデル共にepoch数の増加に従ってlossが減少していることがわかる

両モデル共に過学習の傾向が学習曲線に表れている

各モデルの評価

各モデル・各損失関数でのtestデータでのSI-SNRの平均

Model	SI-SNR
DPRNN 2-mixed (alpha=0.0)	12.59
DPRNN 2-mixed (alpha=1.0)	11.61
DPRNN 3-mixed (alpha=0.0)	4.75
DPRNN 3-mixed (alpha=0.5)	3.74
Conv-TasNet 2-mixed (alpha=0.0)	11.01
Conv-TasNet 2-mixed (alpha=1.0)	10.81
Conv-TasNet 3-mixed (alpha=0.0)	4.43
Conv-TasNet 3-mixed (alpha=0.5)	2.70

modelの評価方法

二つのモデルを二種類の
損失関数で学習させた
各モデルでtestデータをmixさせた音
源の分離を行い、mix前の音源と
比較しSI-SNRを計算した

※alpha=0は従来の損失関数と同義

modelの評価

従来の損失関数を使用したほうが
分離性能(SI-SNR)が高くなってしまう

DPRNNのほうが分離性能が良い

各モデルの評価

各モデル・各損失関数でのtestデータでのSI-SNRの平均

Model	SI-SNR
DPRNN 2-mixed (alpha=0.0)	12.59
DPRNN 2-mixed (alpha=1.0)	11.61
DPRNN 3-mixed (alpha=0.0)	4.75
DPRNN 3-mixed (alpha=0.5)	3.74
Conv-TasNet 2-mixed (alpha=0.0)	11.01
Conv-TasNet 2-mixed (alpha=1.0)	10.81
Conv-TasNet 3-mixed (alpha=0.0)	4.43
Conv-TasNet 3-mixed (alpha=0.5)	2.70

modelの評価方法

二つのモデルを二種類の
損失関数で学習させた
各モデルでtestデータをmixさせた音
源の分離を行い、mix前の音源と
比較しSI-SNRを計算した

※alpha=0は従来の損失関数と同義

modelの評価

従来の損失関数を使用したほうが
分離性能(SI-SNR)が高くなってしまう

DPRNNのほうが分離性能は良い

各モデルの評価

各モデル・各損失関数でのtestデータでのSI-SNRの平均

Model	SI-SNR
DPRNN 2-mixed (alpha=0.0)	**12.59**
DPRNN 2-mixed (alpha=1.0)	11.61
DPRNN 3-mixed (alpha=0.0)	**4.75**
DPRNN 3-mixed (alpha=0.5)	3.74
Conv-TasNet 2-mixed (alpha=0.0)	**11.01**
Conv-TasNet 2-mixed (alpha=1.0)	10.81
Conv-TasNet 3-mixed (alpha=0.0)	**4.43**
Conv-TasNet 3-mixed (alpha=0.5)	2.70

modelの評価方法

二つのモデルを二種類の
損失関数で学習させた
各モデルでtestデータをmixさせた音
源の分離を行い、mix前の音源と
比較しSI-SNRを計算した

※alpha=0は従来の損失関数と同義

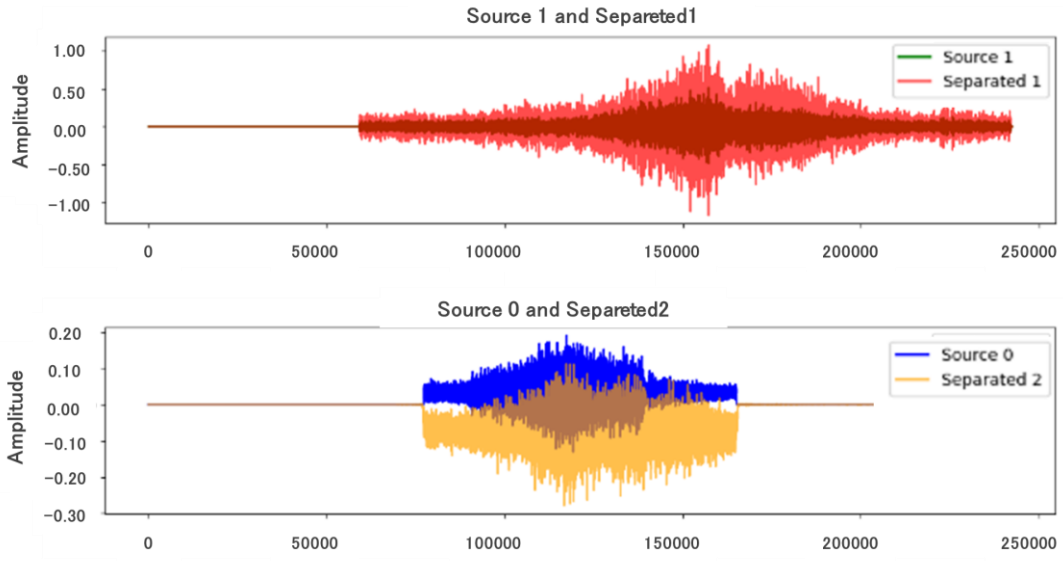
modelの評価

**従来の損失関数を使用したほうが
分離性能(SI-SNR)が高くなってしまう**

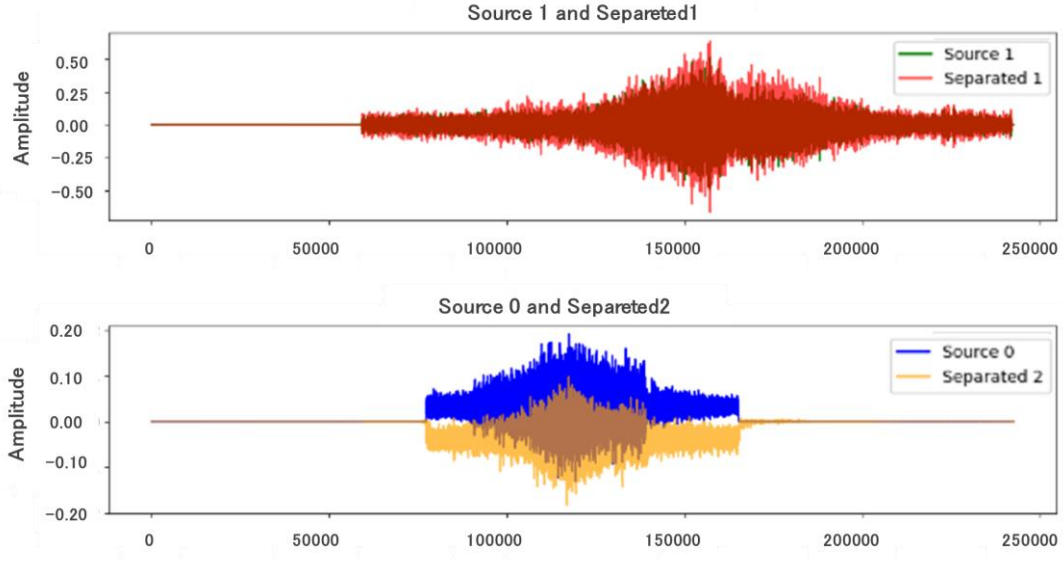
DPRNNのほうが分離性能は高い

音源分離の結果

損失関数ごとの分離結果の比較



DPRNN 2-mixed (alpha=0.0)



DPRNN 2-mixed (alpha=1.0)

DPRNNでそれぞれの損失関数を用い混合した音源を分離し、混合前後の音声を比較する

従来の損失関数に比べて格段に新規の損失関数は音圧の損失がほぼない

音圧の損失が少ないことを優先するため新規の損失関数を使用する

各モデルの評価

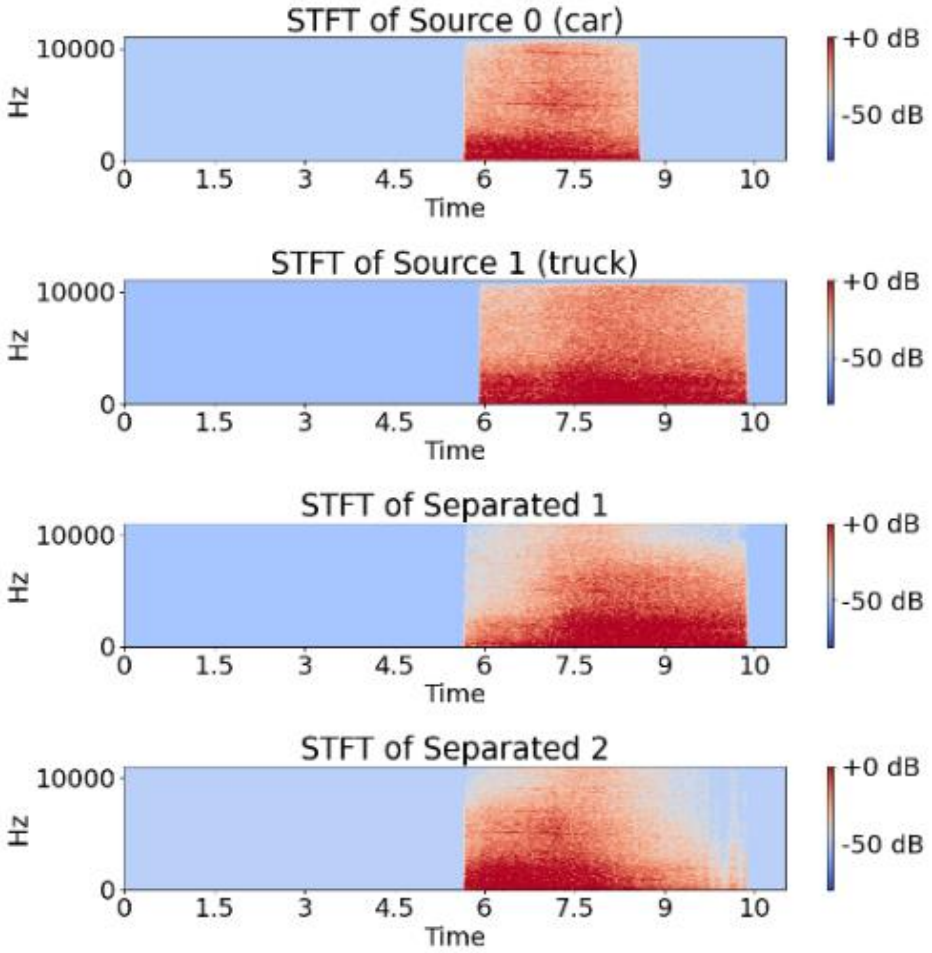
各モデルのtestデータでのSI-SNRの平均

	Conv-TasNet		DPRNN	
	2-mixed	**3-mixed**	**2-mixed**	**3-mixed**
N=32	8.73	2.70	7.33	0.83
N=64	10.81	2.70	11.61	3.74
N=128	11.21	3.23	12.70	4.17

N(モデルの層の深さを決定するハイパーパラメータ)を調整する
モデルの層を深くするほどDPRNNがConv-TasNetの分離性能を上回る結果となった
※N=128以上の数値は研究室の計算機の負荷を超えてしまうが、DPRNNはまだ分離性能が高まる余地がある

考 察

考 察



損失関数ごとの分離結果の比較

分離性能の考察

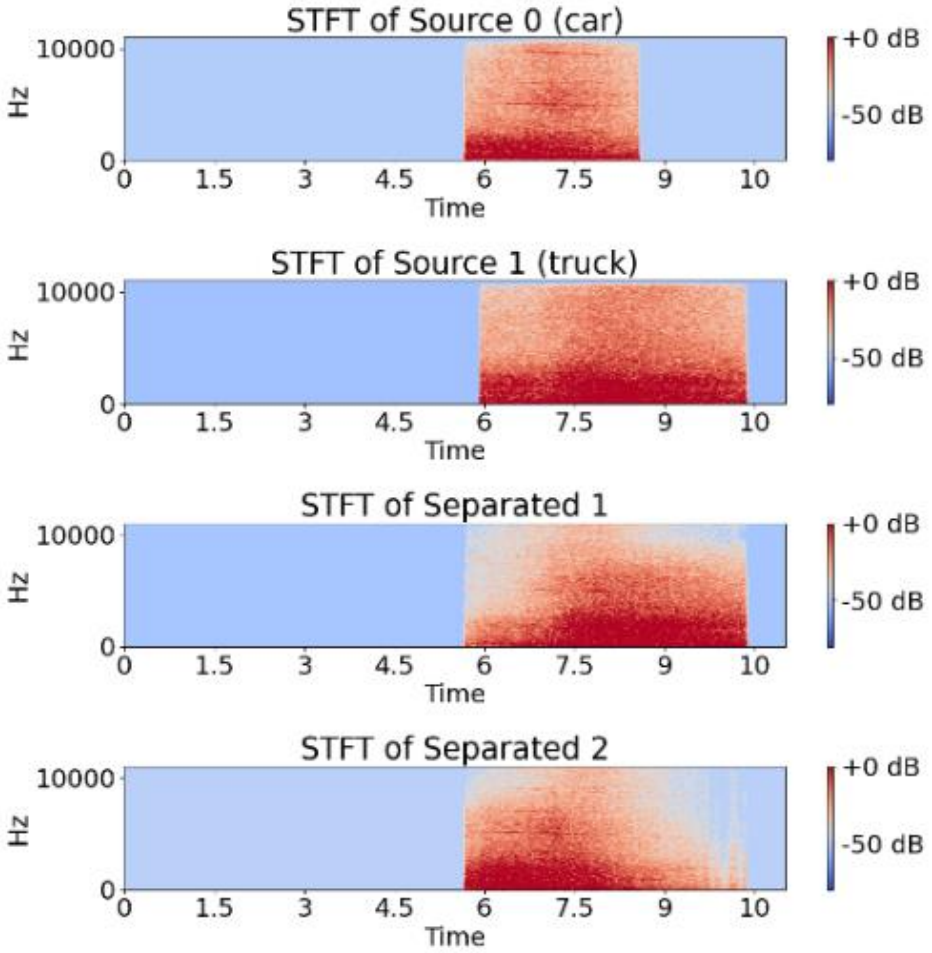
実際に収録した単体車両道路交通騒音のデータセットに対して、DPRNNはConv-TasNetを上回る分離性能を持つことが確認された

過学習への考察

タイヤ路面騒音において各車種が共通する周波数特性を持つことが原因であると考える。

スペクトログラムを見ると非常にランダム性の高い高周波成分をすべての音源が持っている。人間の目でも識別が難しい特性を、音圧ベースの分離モデルで分離することは難しい。

考 察



損失関数ごとの分離結果の比較

分離性能の考察

実際に収録した単体車両道路交通騒音のデータセットに対して、DPRNNはConv-TasNetを上回る分離性能を持つことが確認された

過学習への考察

タイヤ路面騒音において各車種が共通する周波数特性を持つことが原因であると考える。

スペクトログラムを見ると非常にランダム性の高い高周波成分をすべての音源が持っている。人間の目でも識別が難しい特性を、音圧ベースの分離モデルで分離することは難しい。

考 察

過学習の防止

高周波領域の分離が難しい点。
学習データの不足・データポイントの不足が原因と考えられる。
0パディング部分で最適化のリセットがかかっている可能性がある。
ハイパーパラメータの最適化。

↓

ひとつずつ要因を処理することで原因の究明・解決を図る

まとめ

まとめ

研究背景・目的

車両情報を抽出する道路交通情報取得システムの開発

既存の分離モデルを超える精度のモデルを開発する

実験方法

Conv-TasNetとDPRNNで同一の混合車両走行音源
を分離してそれぞれの損失を比較する

結果・考察

長期記憶と入力の圧縮性に優れるDPRNNにより損失が減少
話者分離モデルは騒音分離の分野においても有用である

今後の展望

データセットの増加・44100Hzのサンプリングによる
情報量の増加等を行い顕著にDPRNNモデルの優位性を示す
周波数ベースモデルの検討を再度行う

追加資料

データセット構築

使用データセットの内約

	car	Truck	motorcycle
沼津(実道路)	292	68	54
熊谷(試験場)	68	33	45
total	360	101	99

データセットは沼津市内の道路と、熊谷の交通安全研究所で
録音した単体走行音を使用

今回は44100Hzでの学習を行うと研究室の計算資源では学習が停止してしまうため、
全てのデータを22000Hzにリサンプリングを行う

データセット構築

使用データセットの内約

	car	Truck	motorcycle
沼津(実道路)	292	68	54
熊谷(試験場)	68	33	45
total	360	101	99

データセットは沼津市内の道路と、熊谷の交通安全研究所で
録音した単体走行音を使用

今回は44100Hzでの学習を行うと研究室の計算資源では学習が停止してしまうため、
全てのデータを22000Hzにリサンプリングを行う

研究の妥協点

サンプリング数	22000 Hz
データ数	560 サンプル
層の深さ	最大 128
バッチサイズ	2

データセットは沼津市内の道路と、熊谷の交通安全研究所で
本年度からデータの収集を始めた

DPRNNは計算量が多く、ハイパーパラメータやサンプリング数の増加
することによる計算量の負荷に研究室の計算機では対応することができない

損失関数

$$SI-SNR\_SP = -SI-SNR + alpha * volume\_loss$$

本研究で使用する損失関数

従来の話者分離で使用されていた損失関数

従来の損失関数SI-SNRでは音圧の損失を学習することができないため、

音圧の損失を考慮した損失関数SI-SNR_SPを提案・使用する

損失関数

従来の損失関数SI-SNR(音圧を考慮していない)

$$s_{target} = \frac{<\hat{s},s>s}{\|s\|^2},\ e_{noise} = \hat{s} - s_{target}$$

$$SI-SNR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}$$

※$\hat{S}$：分離後の音源

S：分離前の音源

特にe_{noise}の値が大きくなるほどSI-SNRが小さくなるため

分離後の音源$\hat{S}$を元の音源よりも音圧がかなり大きくなる
ように学習する傾向がある

損失関数

新たな損失関数SI-SNR(音圧を考慮している)

$$SI-SNR\_SP = -SI-SNR + alpha * volume\_loss$$

$$volume\_loss = \log_{10}\left|\frac{||S||^2 - ||\hat{s}||^2}{2e^{-5}}\right|$$

※$\hat{S}$: 分離後の音源

S : 分離前の音源

volume_lossにより音圧を考慮し、分離後の音圧に差があるとき損失が大きくなる

alphaの値を調整しvolume_lossの影響度を制限する。これが大きいと損失が収束しない。

これにより音圧の損失を防ぐことが可能になる

損失関数

$\hat{s}_1 = s + e$

$\hat{s}_2 = 100s + e$の場合を考える(eはノイズ)

$$s_{1\,target} = \frac{<||s||^2 + <e,s>>s}{\|s\|^2} = s + \alpha$$

$$s_{1\,target} = \frac{<||100s||^2 + <e,s>>s}{\|s\|^2} = 100s + \alpha$$

$$\frac{<e,s>s}{\|s\|^2} = \alpha$$

$$e_{1\,noise} = e - \alpha$$

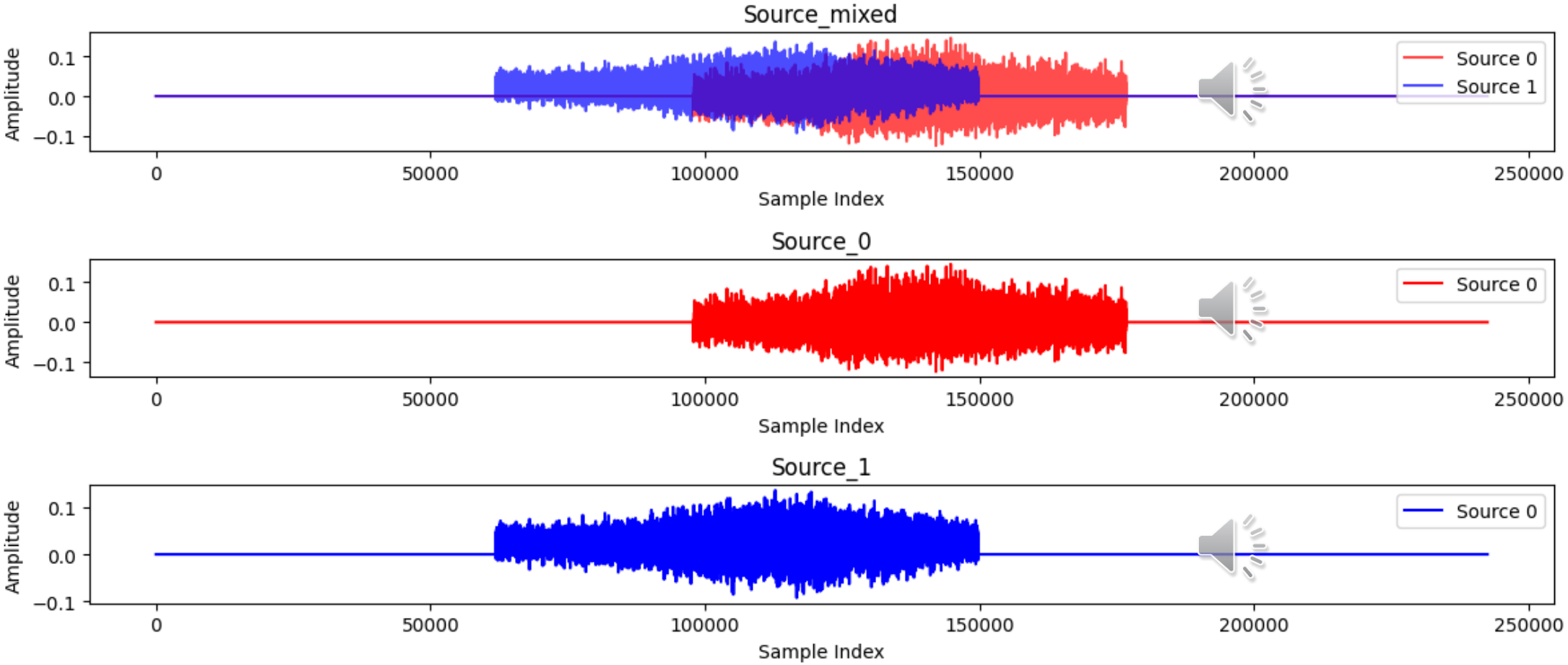
$$e_{2\,noise} = e - \alpha$$

$$s_{target} = \frac{<\hat{s},s>s}{\|s\|^2},\ e_{noise} = \hat{s} - s_{target}$$

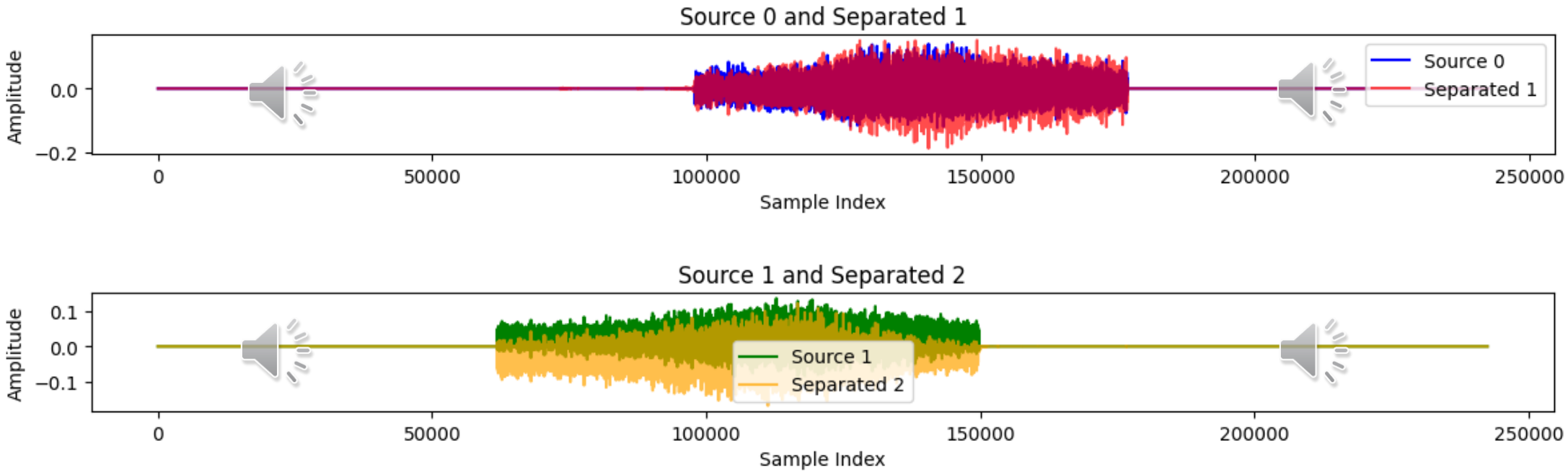
$$SI - SNR = 10\log_{10}\left|\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}\right|$$

※$\hat{s}$：分離後の音源　s：分離前の音源

サンプル音源

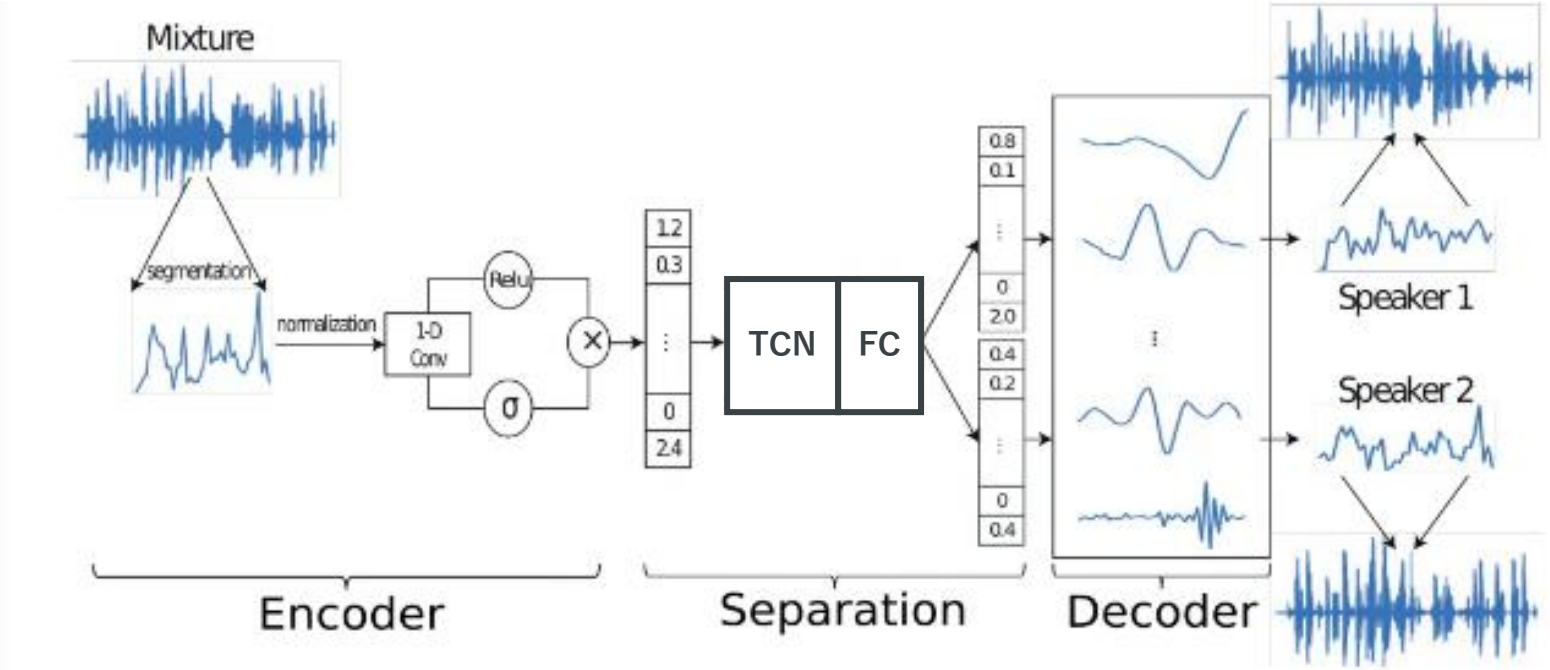

Source_mixed
Amplitude
0.1
0.0
−0.1
0
50000
100000
150000
200000
250000
Sample Index
Source 0
Source 1
Source_0
Amplitude
0.1
0.0
−0.1
0
50000
100000
150000
200000
250000
Sample Index
Source 0
Source_1
Amplitude
0.1
0.0
−0.1
0
50000
100000
150000
200000
250000
Sample Index
Source 0

サンプル音源


Source 0 and Separated 1
Amplitude
0.0
−0.2
0
50000
100000
150000
200000
250000
Sample Index
Source 0
Separated 1
Source 1 and Separated 2
Amplitude
0.1
0.0
−0.1
0
50000
100000
150000
200000
250000
Sample Index
Source 1
Separated 2

採用モデルの構造

Conv-TasNet



エンコーダ

1. 混合音をn個に分割する
2. 抽出した区間を正規化
3. 1DConvにより特徴量抽出
4. 活性化関数を通す

セパレータ

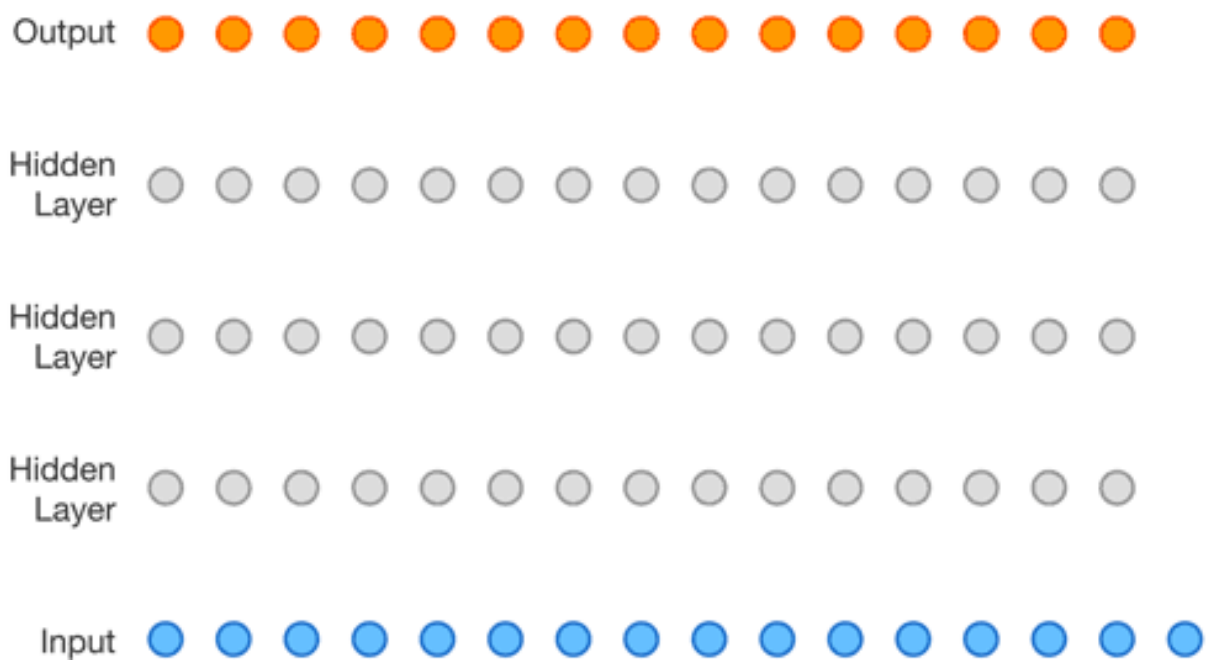
1. TCN(多層1DConv)で特徴抽出
2. 全結合層に特徴量を通し
   混合音源※を分離

※単位波形と一次結合すると元の波形に近似された値が求まる

デコーダ

1. セパレータによって分けられた行列と単位波形の集合を一次結合することで単体騒音を復元

TCN(Temporal Convolutional Network)（補足）



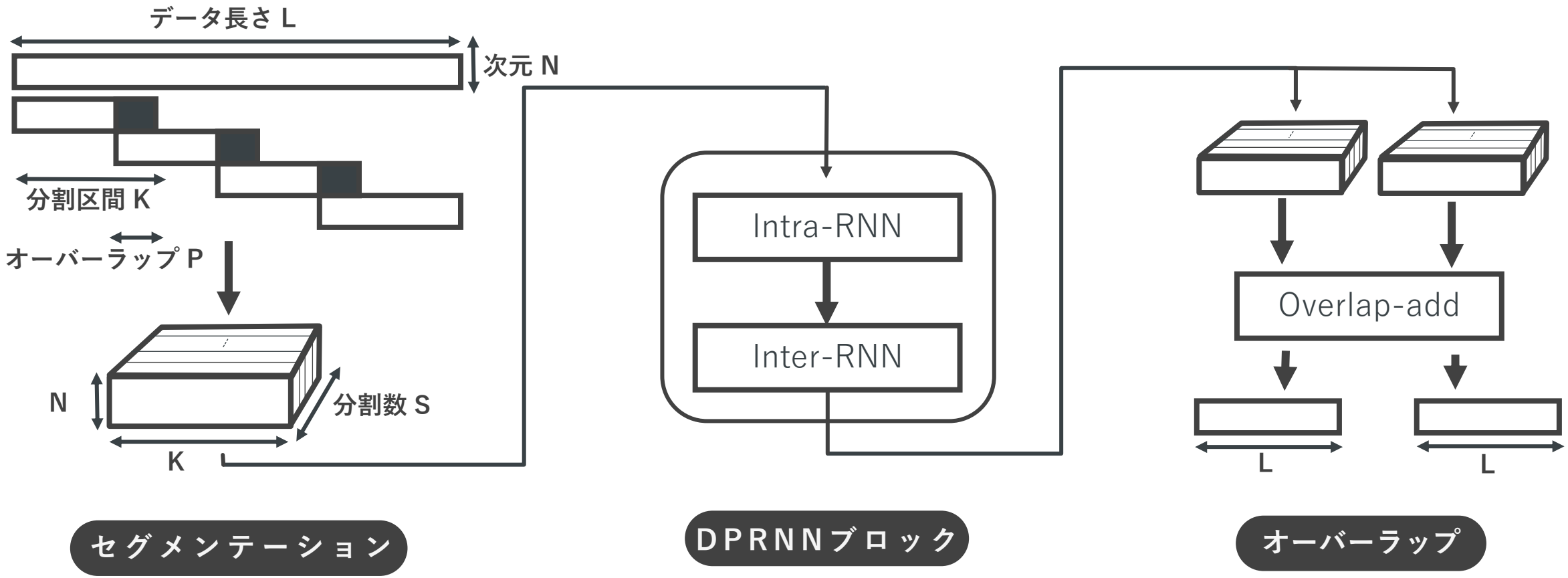
WaveNet: A generative model for raw audio

単にdilationが大きい畳み込み層を1層用いるだけでは，そこまで効力がなさそうに見えるが，層が深くなるにつれて**dilationを大きくすることで，より広範囲の時系列データを反映できるようになる**

⇒層の数をある程度大きくする必要がある

⇒近年は残差パスを導入

Dual path RNN

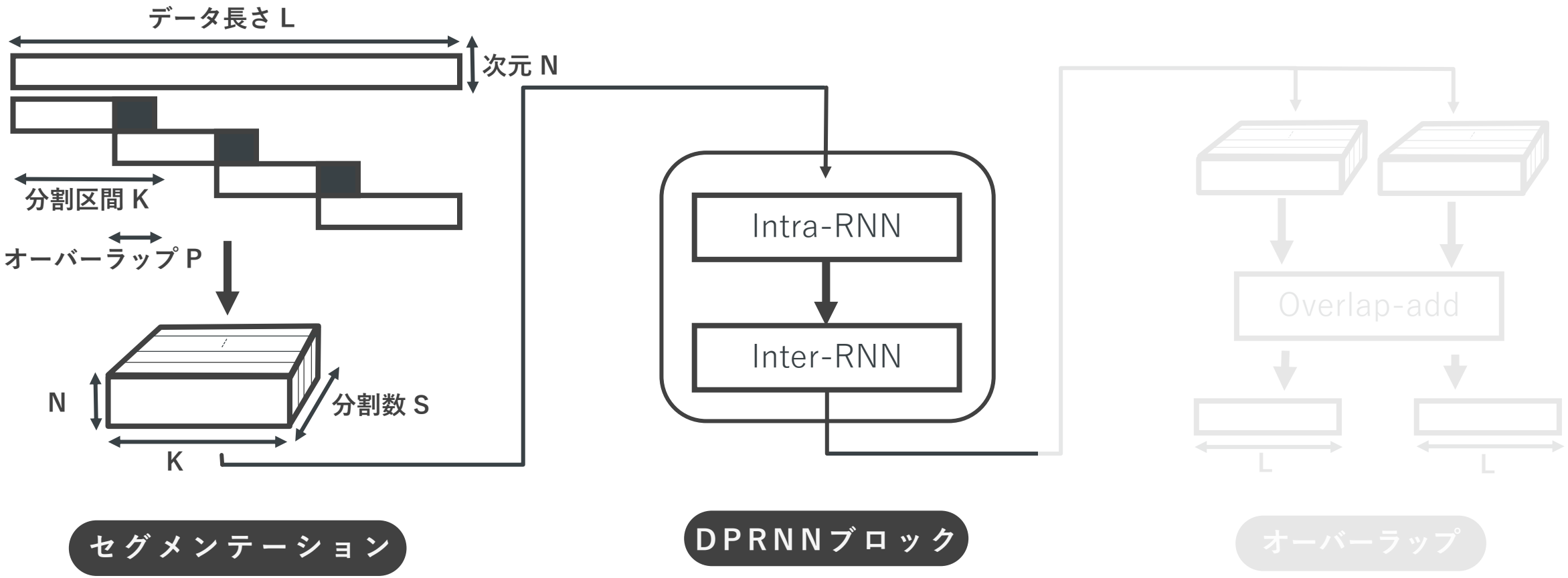


Dual Path RNN(DPRNN)は以下の合計三層の

セグメンテーション層・DPRNNブロック・オーバーラップ層で構築される

今回は特に重要なセグメンテーション層・DPRNNブロックについて説明

Dual path RNN

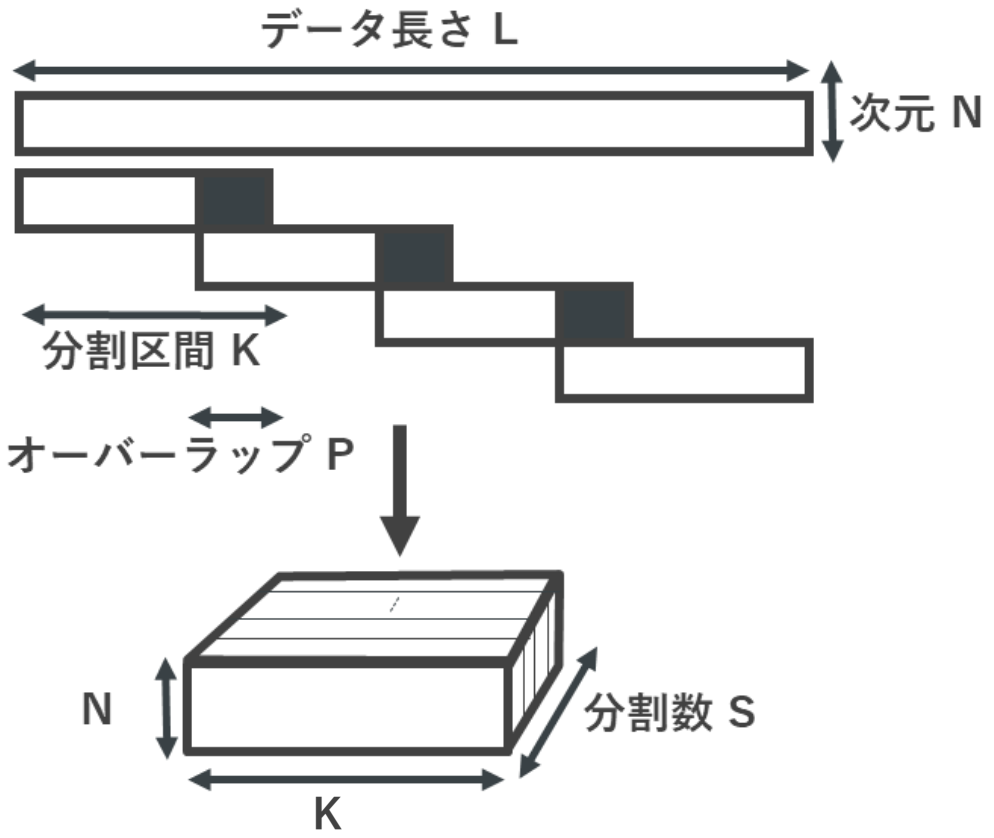


Dual Path RNN(DPRNN)は以下の合計三層の

セグメンテーション層・DPRNNブロック・オーバーラップ層で構築される

今回は特に重要なセグメンテーション層・DPRNNブロックについて説明

セグメンテーション



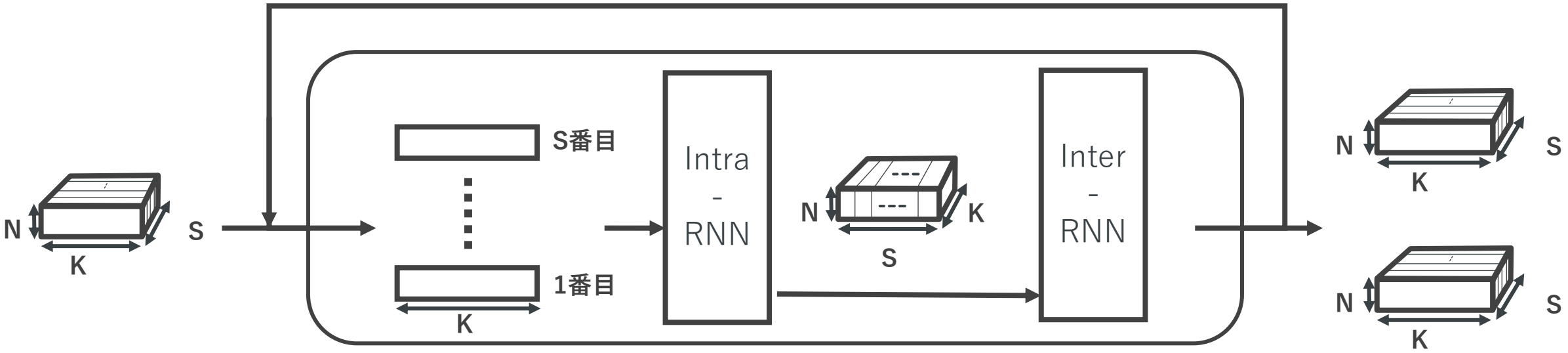
DPRNNブロックへの入力

セグメンテーシ層では以下の手順で
入力データを分割する。

①. シーケンス長さL、次元Nのデータを区間KでS個に分割する
②. この時、区間Kは隣とオーバーラップP区間重なる
③. 分割した区間Kを縦にS個結合し入力テンソルを形成

DPRNNへの入力の長さを $K + S = K + (L/K)$
入力の長さ $K + S$ が最小になるのは $K = \sqrt{L}$
に値が近づくとき
この時入力の長さが $2\sqrt{L}$ となり元のシーケンス長さの平方根に比例するため、ロングシーケンスの圧縮に非常に有用である

セパレート層

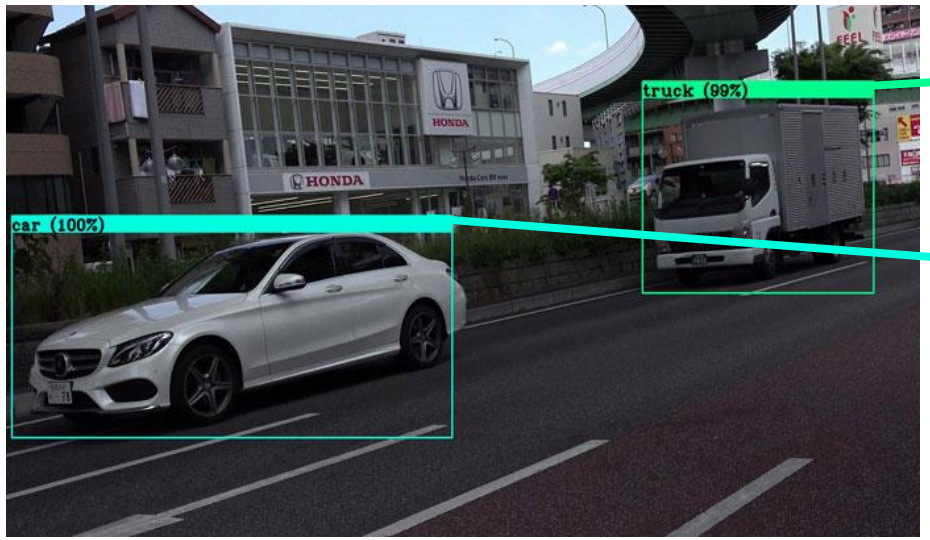


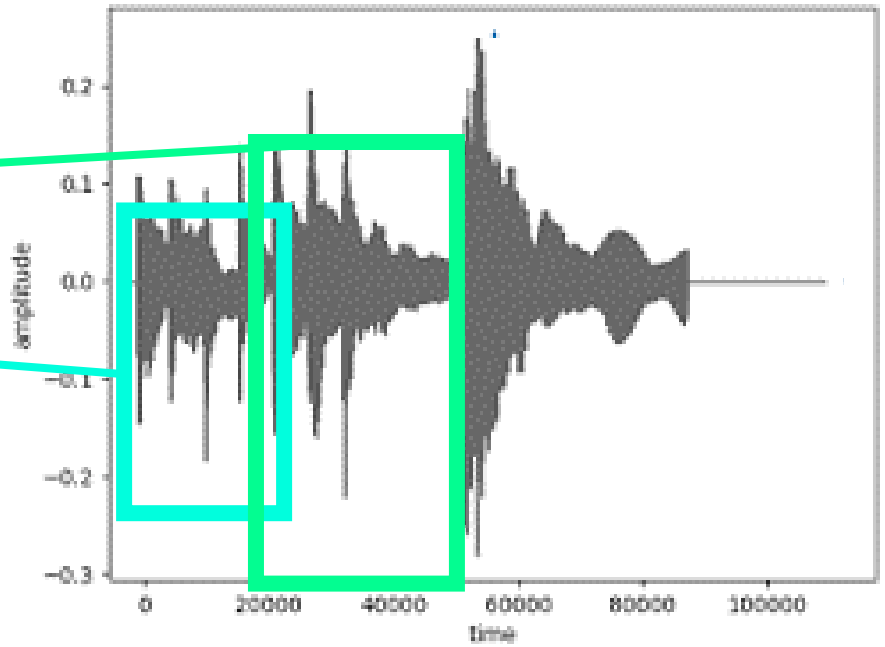
DPRNNの入力　　DPRNNブロック　　DPRNNの出力

・DPRNN層はintra-RNNとinter-RNNが内包された、DPRNNブロックをいくつか直列に接続したもの
・Intra-RNNとinter-RNNの中身は双方向LSTMと全結合層で構成される同じモデルであるが役割が異なる
・Intra-RNNは短期間の波形の特徴を抽出するモデルである。
・入力テンソルを一番目からs番目まで順に一つずつ入れて区切られた区間の波形の特徴量を抽出
・その後iner-RNNでは一番目からs番目までの区間の波形から特徴を抽出したものをすべて連結したものを
・モデルに通し波形全体の特徴量を求める、データは大きいけどシーケンスはKのまま
・これを繰り返し、最終的にそれぞれの入力波形のもとになる波形を復元する

データセット構築

データセットは実走行音を録画し音源のみを使用した



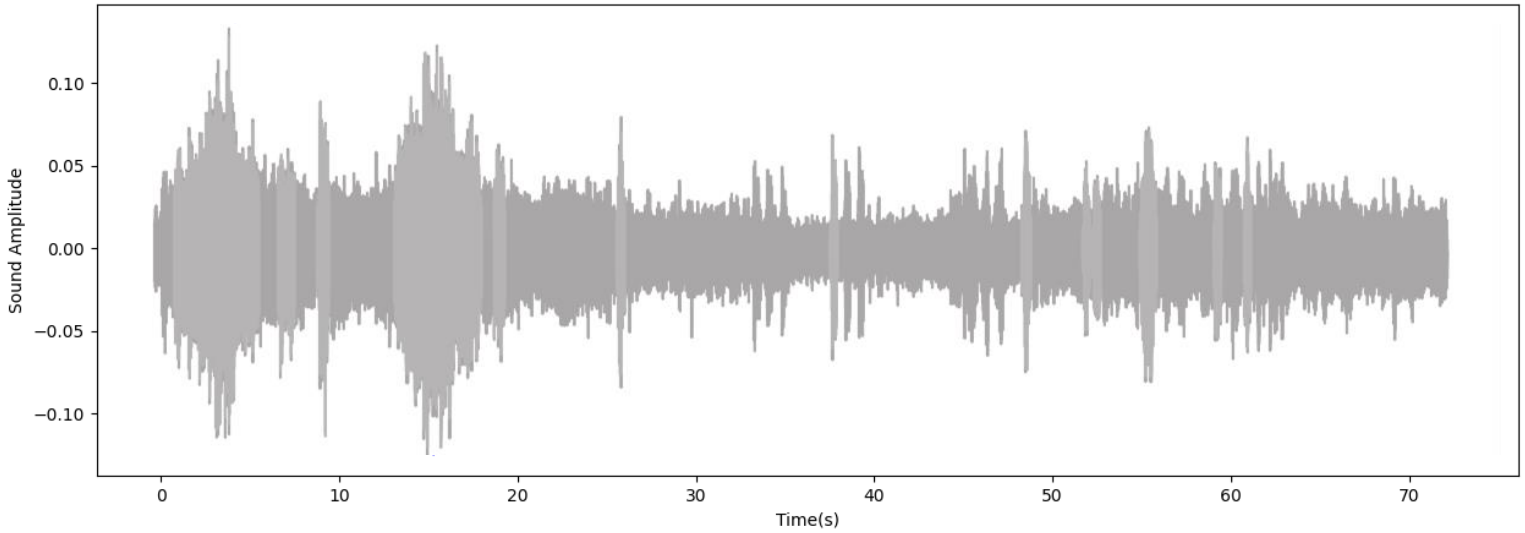


録画データをmp3に変換したもの

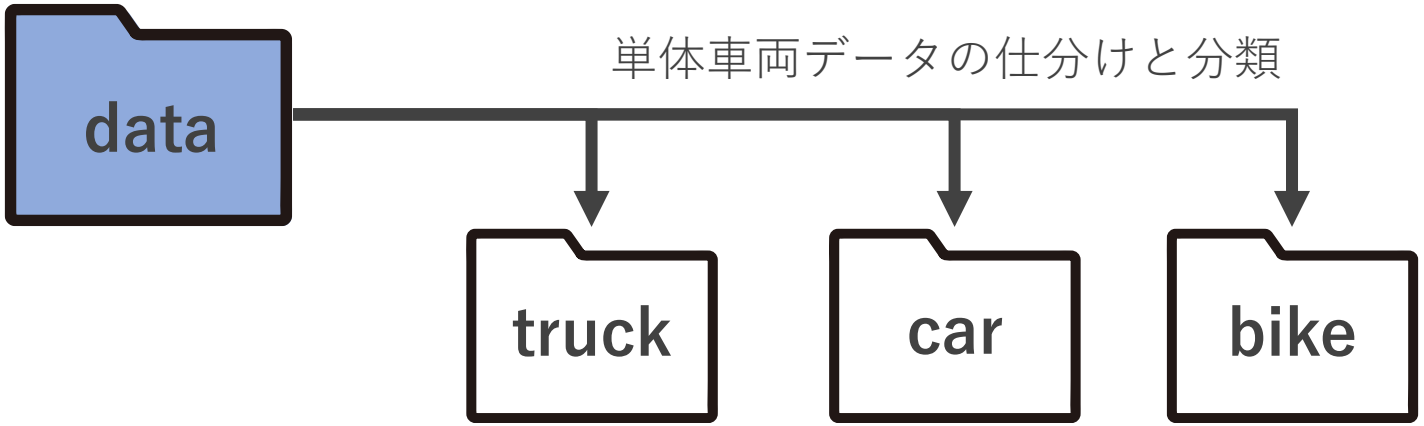
YOLOv7を用いて車両が一台のみ通過しているフレームを
指定して動画を切り抜きデータセットを作成

データセット構築

一定以上以上の音圧区間を切り抜いたデータ



単体車両データの仕分けと分類



Step 1

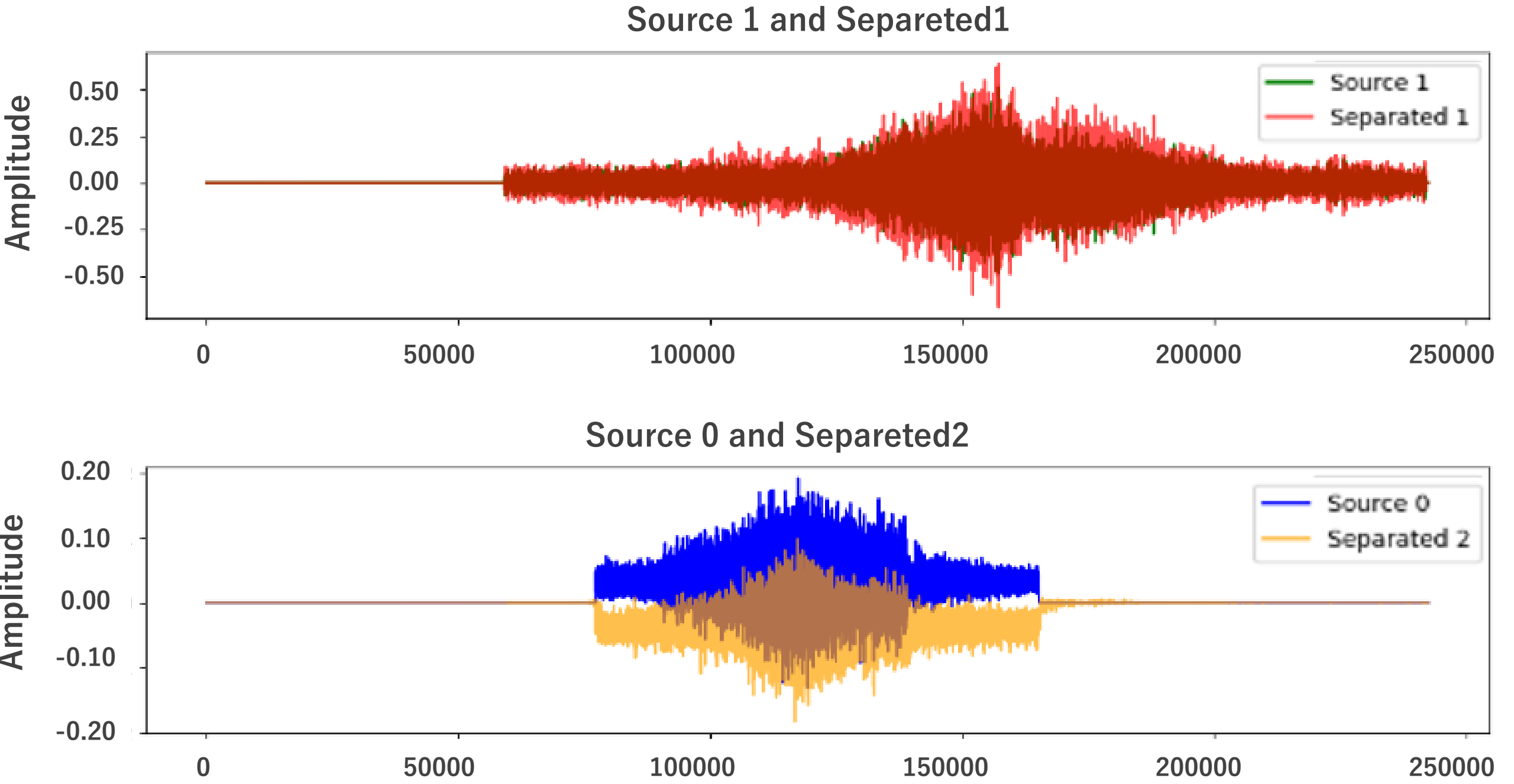
動画を一定以上の音圧の
区間でトリミングする

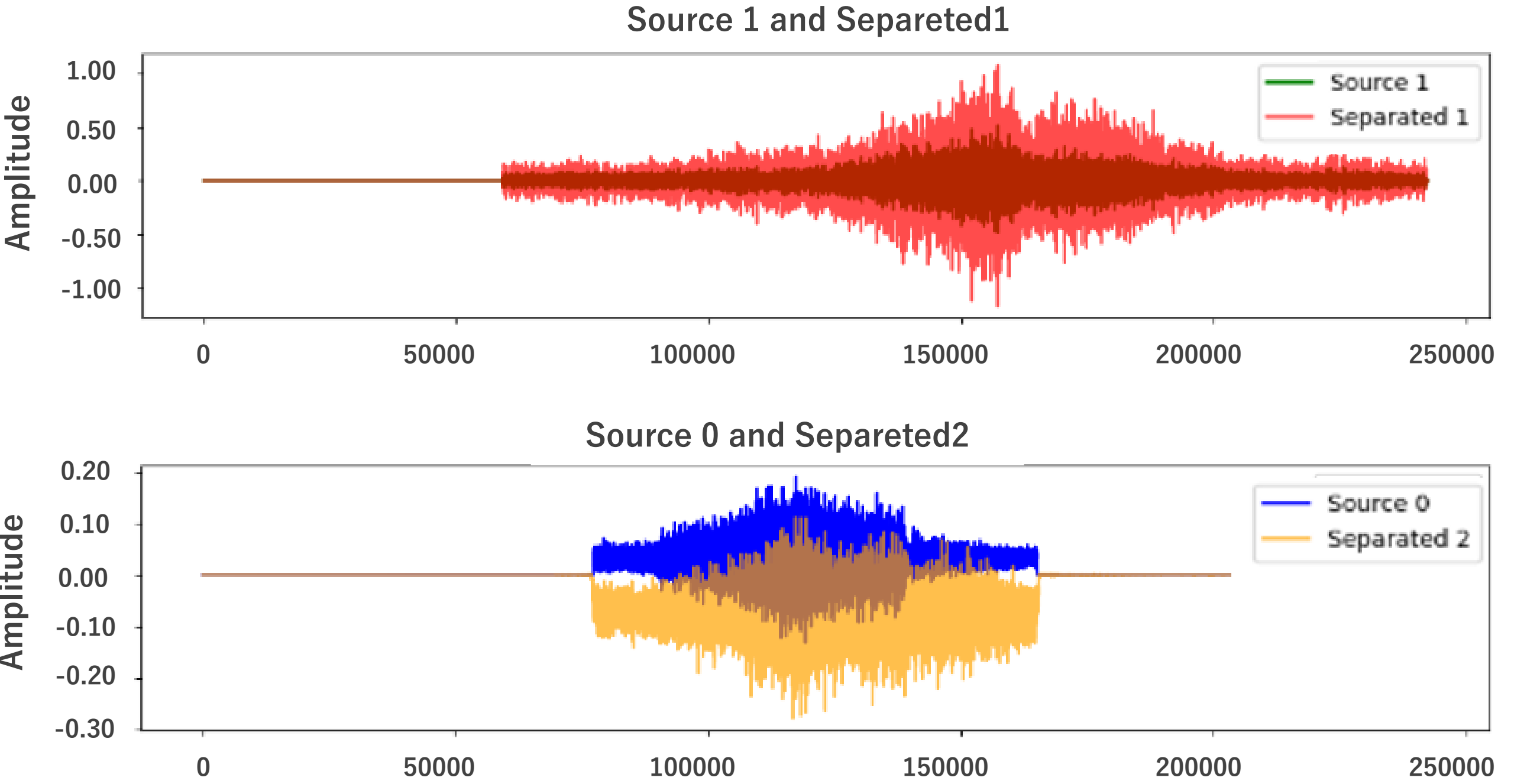
Step 2

トリミングした区間をＹOLOv7
を用いて単体車両のみかを確認

Step 3

単一車両データを車種
ごとに分類する


Source 1 and Separated1
Amplitude
0.50
0.25
0.00
-0.25
-0.50
Source 1
Separated 1
0
50000
100000
150000
200000
250000
Source 0 and Separated2
Amplitude
0.20
0.10
0.00
-0.10
-0.20
Source 0
Separated 2
0
50000
100000
150000
200000
250000


Source 1 and Separated1
Amplitude
1.00
0.50
0.00
-0.50
-1.00
Source 1
Separated 1
0
50000
100000
150000
200000
250000
Source 0 and Separated2
Amplitude
0.20
0.10
0.00
-0.10
-0.20
-0.30
Source 0
Separated 2
0
50000
100000
150000
200000
250000

memo

結果が出なかった場合
機械の振動や騒音は特徴量や周波数に分けると正規分布をとる定常性を持つ
交通騒音はエンジン駆動音は定常性を持つが、タイヤの路面音は非定常性を持つ
つまり交通騒音は非定常性と正常性を二つ持つ値であり扱いが難しく
非定常音のみの分離を目的として構成されたDPRNNだと不足がある可能性がある

memo

周波数領域のデータは非常に大切で分離に有用だが
位相のデータが欠損してしまう
そのため、現在はムーブメントとして扱いやすい時間領域のモデルが論文を占めてるけど
位相の問題を解決したら周波数領域のモデルがはやると思う

m e m o

今回音声は時間領域データとして扱うので
損失は各フレームごとの音圧の差を取る
当然差の値は正負の値を取るため損失は平均二乗誤差を
用いて表す

memo

人数が増えると爆発的に難易度が高まる
Convtasnetは4人が体感的に限界な気がする
DPRNNはどこまでいけるか
正直max9車両分類したいけどかなりしんどいと思う
あと、どっちのモデルも最初に何種類に分類するかは指定しないといけない
交通をカメラで監視してどの車両か何種類走行しているか確認し、それに合わせたモデルを切り替えるし
入力も与える

スライド1: Conv-TasNetの特徴

Conv-TasNetは、音声分離のためのモデルで、エンコーダ、セパレータ、デコーダの三層構造を持っています。エンコーダでは、入力音声を正規化し、1D-Conv（一次元畳み込み）を用いて波形の特徴量を抽出します。この際、ReLU関数を活性化関数として使用し、負の値を削除することで、波形を単位波形の和で表現します。セパレータでは、エンコーダからの出力をTCN（時系列畳み込みネットワーク）に入力し、混合音源から単体音源を分離します。最後に、デコーダでは、セパレータからの出力と単位波形を結合し、単体音源を復元します。

スライド2: DPRNNの特徴

DPRNNは、音声分離のためのモデルで、セグメンテーション層、セパレート層、オーバーラップ層の三層構造を持っています。セグメンテーション層では、入力音声を一定の区間で分割し、それらを縦に結合して入力テンソルを形成します。この際、区間の長さと分割数は、全体のシーケンス長の平方根に比例するように設定されます。セパレート層では、セグメンテーション層からの出力をDPRNNブロックに入力し、音源を分離します。

DPRNNは、intra-RNNとinter-RNNの二つのモデルで構成され、それぞれが波形の局所的な特徴と全体的な特徴を捉える役割を果たします。それにより、波形の特徴を正確

memo