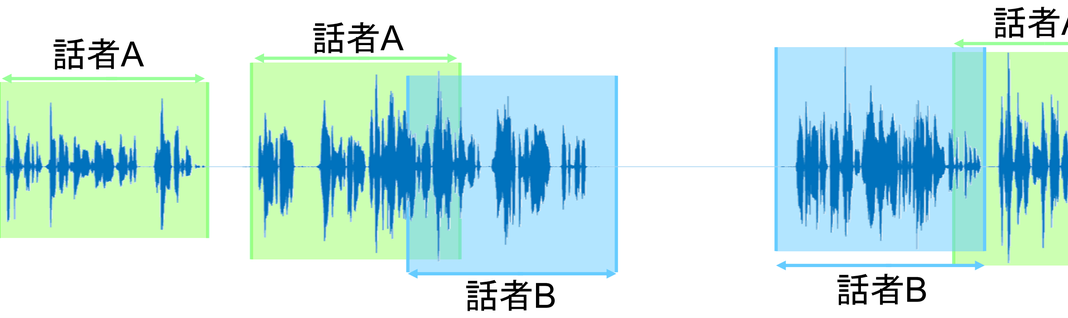
42

ブラインド音源分離のための単一話者発話区間

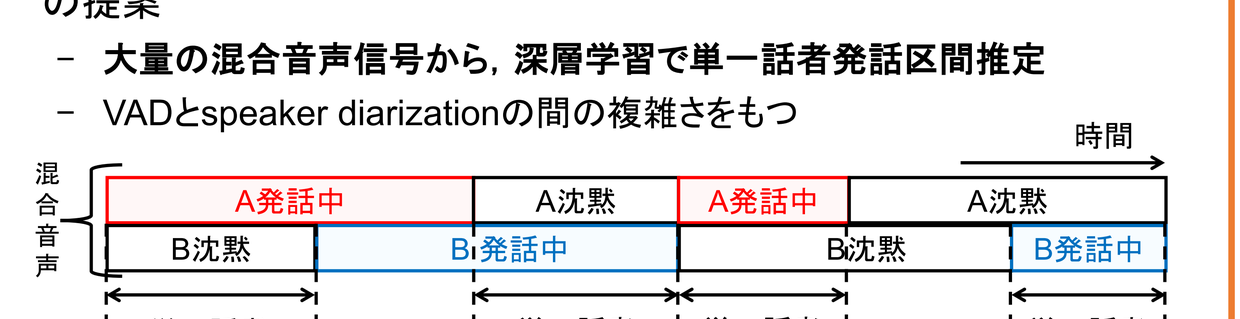
☆加藤大輝，川口翔也，北村大地（香川高専）

1. 研究背景

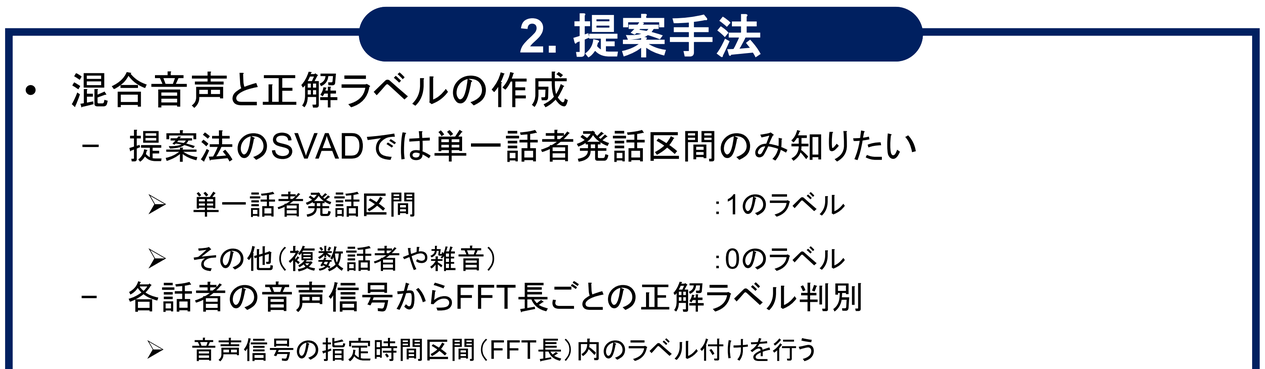
- ブラインド音源分離（blind source separation: BSS）
  - 複数の音源が混ざり合った信号から各音源の信号を推定



- 音声発話検出（voice activity detection: VAD）[Srinivasan+,1993] [Vini+,2021]
  - 音声信号に対して「発話区間」と「非発話区間」を判別



- 話者ダイアライゼーション（speaker diarization）[Park+, 2022] [Ishiguro+, 2012]
  - 音声信号に対してVADに加えて「各話者の発話区間」も判別

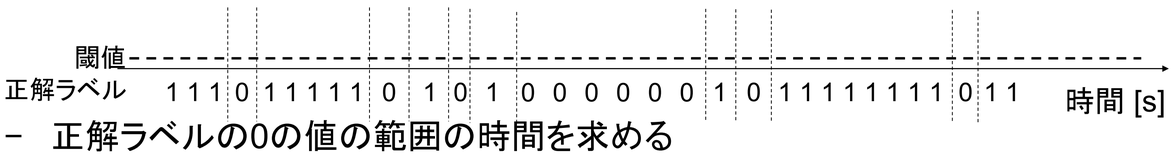


本研究の目的

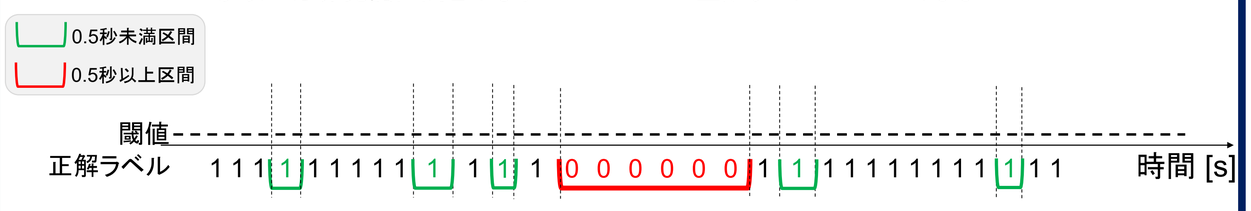
- BSS性能向上のため一話者のみが発話している時間区間が多く必要 [Junkun+, 2023]
- 単一話者発話区間検出（single voice activity detection: SVAD）の提案
  - **大量の混合音声信号から，深層学習で単一話者発話区間推定**
  - VADとspeaker diarizationの間の複雑さをもつ

2. 提案手法

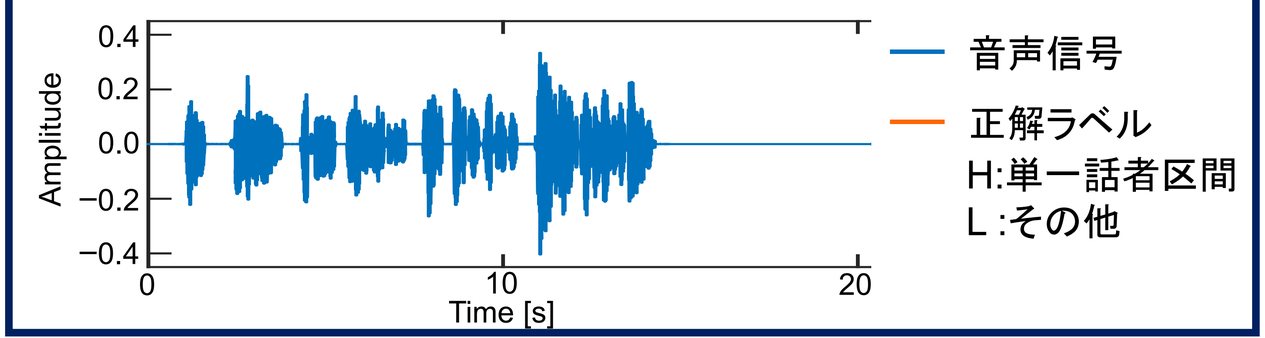
- 混合音声と正解ラベルの作成
  - 提案法のSVADでは単一話者発話区間のみ知りたい
    - 単一話者発話区間　：1のラベル
    - その他（複数話者や雑音）　：0のラベル
  - 各話者の音声信号からFFT長ごとの正解ラベル判別
    - 音声信号の指定時間区間（FFT長）内のラベル付けを行う
    - 音声信号の絶対値をとり，単一話者区間判別用の音量の閾値を設定（発話/雑音の判別）



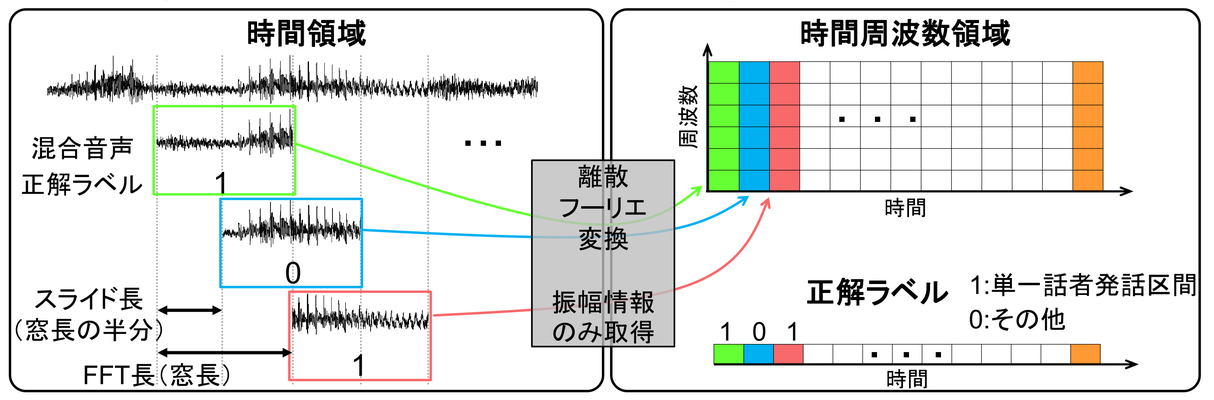
  - 正解ラベルの0の値の範囲の時間を求める
    - 0.5秒未満は連続音声とみなす　正解ラベル　0→1に変更
    - 0.5秒以上は非発話区間とみなす　正解ラベル　0→0のまま



    - それぞれの正解ラベルの和を取り，1になる区間を単一話者発話区間として1のラベル付けをそれ以外を非単一話者区間とし0のラベル付けを行う

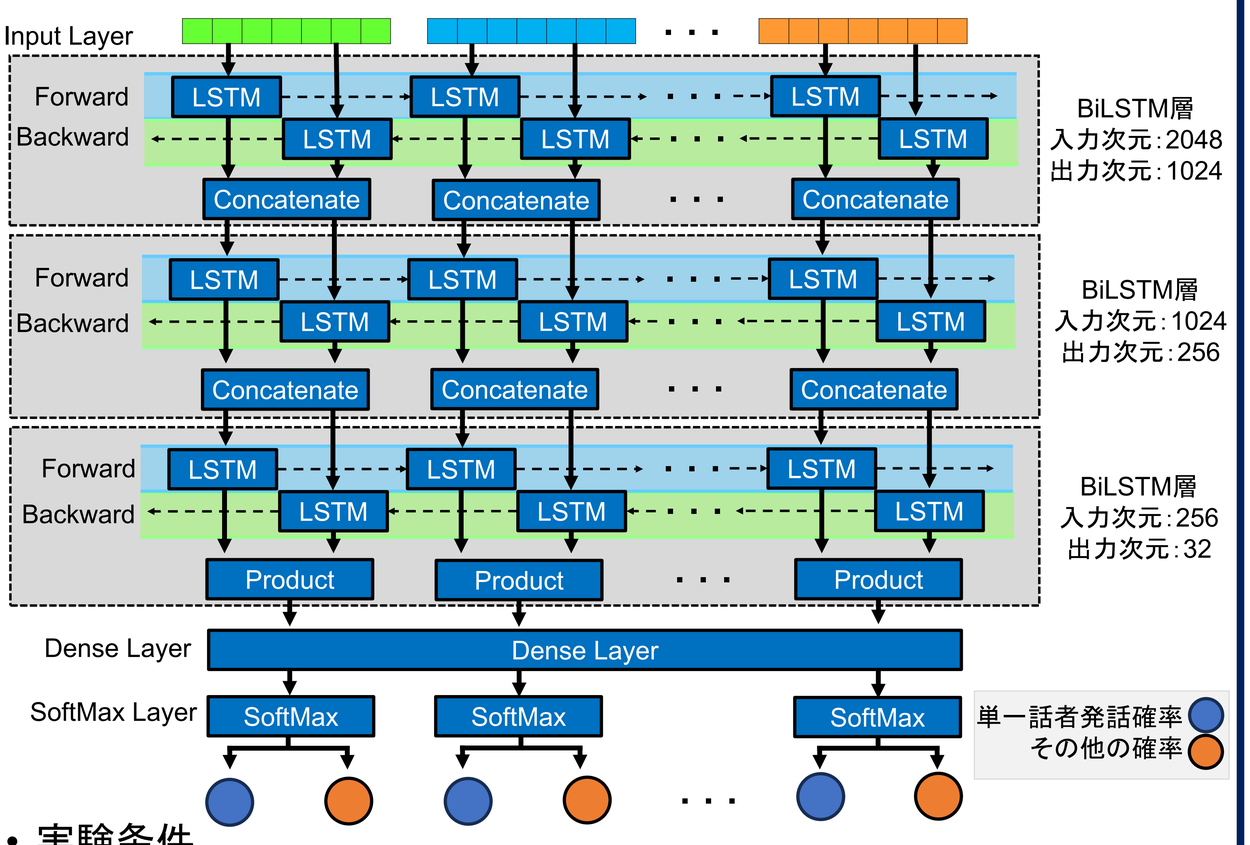


- 深層ニューラルネットワークに入力するデータの前処理
  - FFT長ごとに振幅スペクトログラムの作成を行う



3. 実験

- BiLSTMを用いた深層ニューラルネットワークを使用し，混合音声振幅スペクトログラムから単一話者発話区間の推定をおこなう

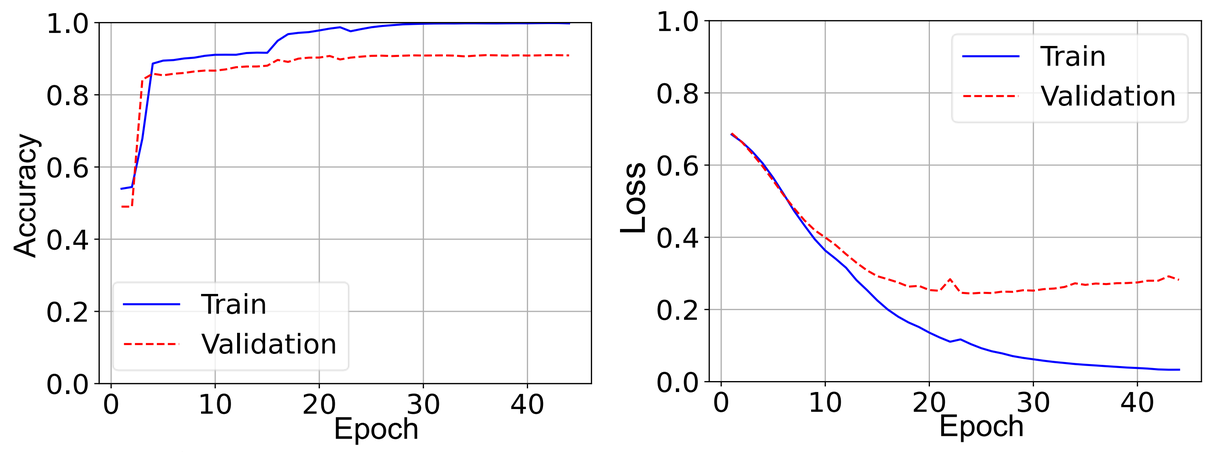


- 実験条件

使用する音声信号	公開データセットJapanese versatile speech（JVE）corpus [Takamichi +,2019]
学習用音声信号	2人の混合音声信号（60秒）を60波形,8組分作成
重み探索用検証データ	2人の混合音声信号（60秒）を60波形,5組分作成
テストデータ	2人の混合音声信号（60秒）を60波形,5組分作成
学習率	1e-4
サンプリング周波数	16 [kHz]
FFT長	256 [ms]
スライド長	128 [ms]

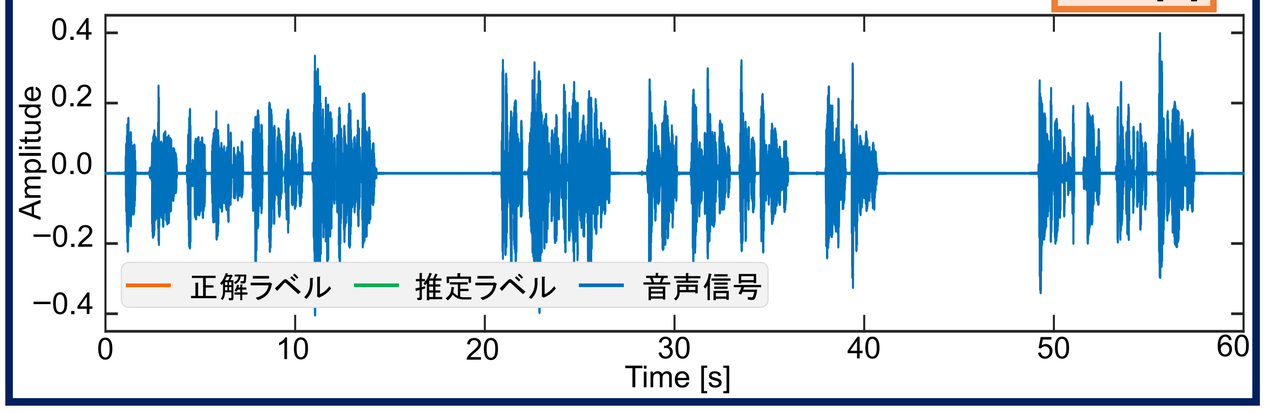
4. 実験結果

- 学習データと検証データにおけるLossとAccuracy



- テストデータにおける推定ラベルと正解ラベルの比較

正解率
92.52 [%]



本研究のまとめ

- 推定ラベルと正解ラベルの比較より次のことがわかる
  - 殆どの時間区間で正確な推定ができているが，沈黙区間が1~2秒の発話区間推定が正確にできていない
  - 発話区間推定の開始，終了地点に数百[ms]のずれが生じている