wav2vecモデルによる方言音声資料のテキスト化

峯尾海成, LI XIAORAN, 谷口ジョイ, 高野敏明
静岡理工科大学

2022

言語資源

(1) 方言学者である故・山口幸洋氏の遺した約2,000点の音声資料
(2) 1950年以降に全国各地で録音された自然談話資料
(3) 記録媒体：オープンリール，DAT，カセットテープ，MD等

本言語資源の利点・問題点

【利点】 当時の方言を知る上で非常に貴重であり、既に消失された言語的な特徴を知るのに有用
【問題点】 データ量が膨大である上，音質が劣悪であることが多い
書き起こしに多くの時間，労力を要する

問題の解決方法提案

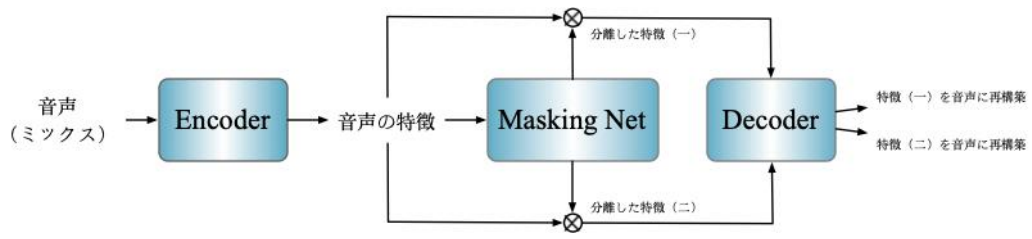
- 言語資源の文字化における自動音声認識の活用
- SepFormerモデルによる話者の分離
  → 本発表では，試験的に，1955年及び1965年に録音された静岡井川方言のデータを用いる
- 旧井川村は、長く周囲から隔絶されており、周辺とは形態、統語的に大きく異なる。人口減少が進み、井川方言の話者は200名以下と推測される。

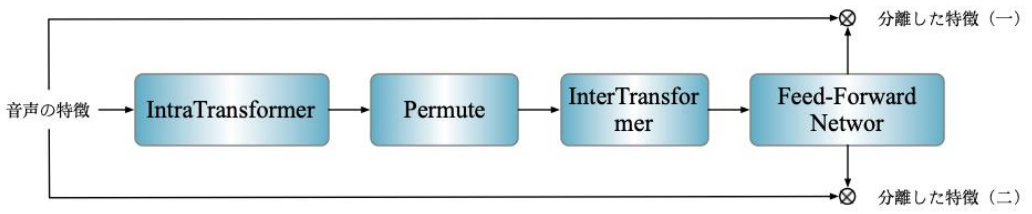
実際の音声資料（サンプル）

話者分離モデル・音声認識モデル

SepFormerモデルによる話者の分離

【目的】複数の話者による自然会話は，音声の重なりが多く，書き起こしが困難である。話者ごとに音声を分離したものへと変換することで、音声認識の精度を向上させることを試みる。



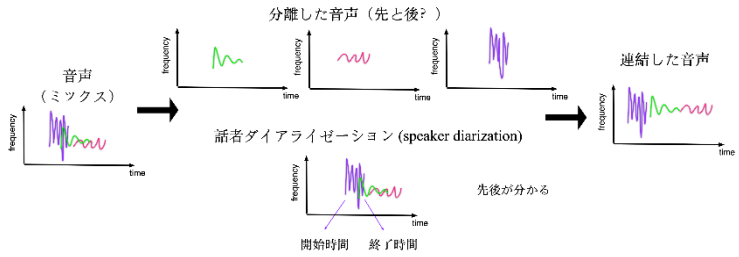


【手法】
(1) 当時の方言を記録した音源はほとんど残されておらず，広範な学習データセットが存在しないため，事前学習（Per-train）した音源分離モデルを用いて，自作の静岡井川方言のテストデータ上で推論（Inference）を行い，音声認識にかける。今回は，WHAMR! データセット[1]で事前学習した SepFormer モデル[2]を採用した。
(2) 訓練プロセスは上図に示す。

[1] 16kサンプリング周波数のWHAMR!データセット[1]（＝WHAMR!はWSJ0-Mixデータセットを高ノイズ化したもの）

話者ダイアライゼーションによる順序判別

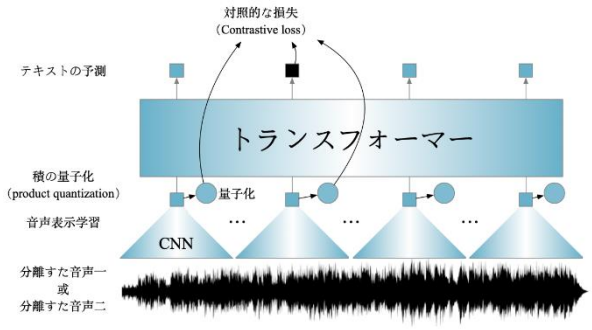
【目的】複数話者の発話順序を判別し、「いつ誰が話したか」を推測する。



【手法】クラスタリング手法（＝データ間の類似度に基づき，データを分類する手法）によって，それぞれの話者の発話の開始・終了時間を推測する「話者ダイアライゼーション（Speaker Diarization）技術[3]」を用いる。

wav2vecモデルによる音声認識

【目的】 複数話者の音声の並べ替えと重ね継ぎ（splicing）を行うことで、環境雑音や残響を伴う複数話者による会話を自動認識する。



【手法】 JSUT日本語データセット[6]で微調整（fine-tune）したFacebookのXLSR-Wav2Vec2モデルを用いる。Wav2Vec2の訓練プロセスは上図に示す。

まとめ

WER（Word Error Rate，単語誤り率） $\frac{\text{挿入単語数+置換単語数+削除単語数}}{\text{正解単語数}}$

CER（Character Error Rate，文字誤り率） $\frac{\text{挿入語数+置換語数+削除語数}}{\text{正解語数}}$

PER（Phoneme Error Rate，音素誤り率） $\frac{\text{挿入音素数+置換音素数+削除音素数}}{\text{正解音素数}}$

	WER	単語数	誤り単語数
T10_raw	0.988399	862	852
T10_concat	0.988349	1030	1018
	CER	文字数	誤り文字数
T10_raw	0.826784	1247	1031
T10_concat	0.924721	1076	995
	PER	音素数	誤り音素数
T10_raw	0.690177	2708	1869
T10_concat	0.892825	2268	2041

プロのトランスクライバーによる文字起こし（一例）[2]

話者2 うんいううんうん眠いからあそうか(＃＃＃＃@00:00:20)
話者1 あったらそう(＃＃＃＃@00:00:26)だんだん(＃＃＃＃@00:00:26)
話者2 そういう終わるものやみだしただったらとかなんだら(＃＃＃＃@00:00:37)
話者1 そうじゃねややっぱりさむいでな
話者2 何度も大きくなってもりも(＃＃＃＃@00:00:47)

[2]「＃」はunclearである音声を示す。

実際の音声資料(t10)
（冒頭60秒程度）

成果及び今後の課題

【成果】
- オリジナルモデルにより評価を行ったところ、WER（単語誤り率）はわずかに減少し、CER（文字誤り率）は上昇し、またPER（音素誤り率）は上昇した。
- CERが上昇した原因として、今回用いた手法が音源分離に依存しており、うまく分離できなかった音声がある場合に、音声が重複してしまうためと推察される。
- CERにおいては、以下のような結果のセンテンスも存在した。

音声認識
猿下て切ったほれぬままの末ませってるだまならでな内勢に狭にさる育員滑メにあってさな言ったたドがあるだやと

人間（正解）
それでちったぁほれにもも荷物もしょってるだもんだでな1日ふんとに大変な目に遭ってその行ったことがあるだよんあ

CER: 0.667

【課題】
- 本研究では話者分離と音声認識の手法を用いたが、音声中に不明瞭な音が多数残存したものについては，改善の余地がある。今後は、微調整（fine tune）により，精度を向上させる必要がある。また、今回は方言に合わせたファインチューニング（音声認識）を行っていないため、今後は方言のイントネーションなどを考慮することが必要である。
- 本研究で用いた手法は比較的新しい手法であり、話者分離の技術を今後方言研究に活用するためには、さらなる検討が必要である。
- 方言の音声認識については、現状では資料やアルゴリズムが少なく、それらの更なる充実が期待される。

謝辞
本研究で扱う言語資源を提供してくださった故・山口幸弘先生のご家族に深謝いたします。

文献及び関連URL
[1] Subakan C, Ravanelli M, Cornell S, et al. Attention is all you need in speech separation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 21-25.
[2] http://wham.whisper.ai/
[3] https://github.com/pyannote/pyannote-audio
[4] https://huggingface.co/facebook/wav2vec2-large-xlsr-53
[5] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-japanese
[6] https://sites.google.com/site/shinnosuketakamichi/publication/jsut
[7] https://qiita.com/Kchan/items/7bba1f066234ba24898b
[8]松浦考平, 三村正人, 河原達也 (2021), アイヌ民話アーカイブに対する音声認識,『自然言語処理』28(3), 824–846. https://doi.org/10.5715/jnlp.28.824