複数の装着型マイクを用いた多人数会話音声認識に関する検討

☆林升柯，　綱川隆司，　西田昌史，　西村雅史（静岡大院・情）

1　はじめに

講義などの多人数会話を音声認識して分析したいという要望があるが，接話マイクを用いても発話重畳による認識性能低下の問題は避けられない．発話重畳の影響を軽減する方法の一つとして，喉の振動を直接拾う咽喉マイクを利用する音声収録方法がある．しかし，咽喉マイクは通常のマイクに比べて収録できる帯域が狭く特性も異なるため，接話マイクなどを想定した通常の音響モデルでは音響ミスマッチによる認識性能低下が生じる．

咽喉マイクを利用した先行研究として我々は認識後に単語信頼度に基づいて認識結果を選択する手法を提案した[1]．しかし，認識後における選択では，選択されなかったマイクの情報は一切活用されない．本論文では音声を接話マイクと咽喉マイクで同時に収録し，あらかじめ DNN で雑音を抑制した音声を推定することで，音声認識率を改善する方法について検討したので報告する．

2　提案法

提案法では接話マイクで収録した音声と雑音の少ない咽喉マイクの音声を併用し、音響ミスマッチが少なく、かつ雑音が少ない音声を、DNN を用いて推定することで認識性能の向上を図る．

図 1 に示すように提案する DNN は変換部と推定部から構成されており，接話マイクと咽喉マイクのスペクトル，及び発話重畳区間のラベルを入力としている．変換部では LSTM によって時系列を考慮しつつ，咽喉マイクのスペクトルを帯域拡張することで，音響ミスマッチを抑制する．推定部では接話マイクと咽喉マイクのスペクトルがそれぞれ 2 層の全結合層で接続された後，マージ層で結合され，さらに 2 層の全結合層で接続されることによりクリーンなスペクトルを推定する．このとき接話マイクの全結合層とマージ層の間はゲートで接続されており，接話マイクの全結合層の出力を制御できるようにしている．発話重畳した際にゲートを絞ることにより，接話マイクに重畳した雑音の影響を軽減する．この結果クリーン環境では接話マイクのスペクトル情報を多く用いて推定が行われ，発話重畳が起きたときは咽喉マイクのスペクトル情報を多く用いて推定が行われる．最終的に推定したスペクトルを、逆短時間フーリエ変換により音声に変換し，音声認識を実行する．

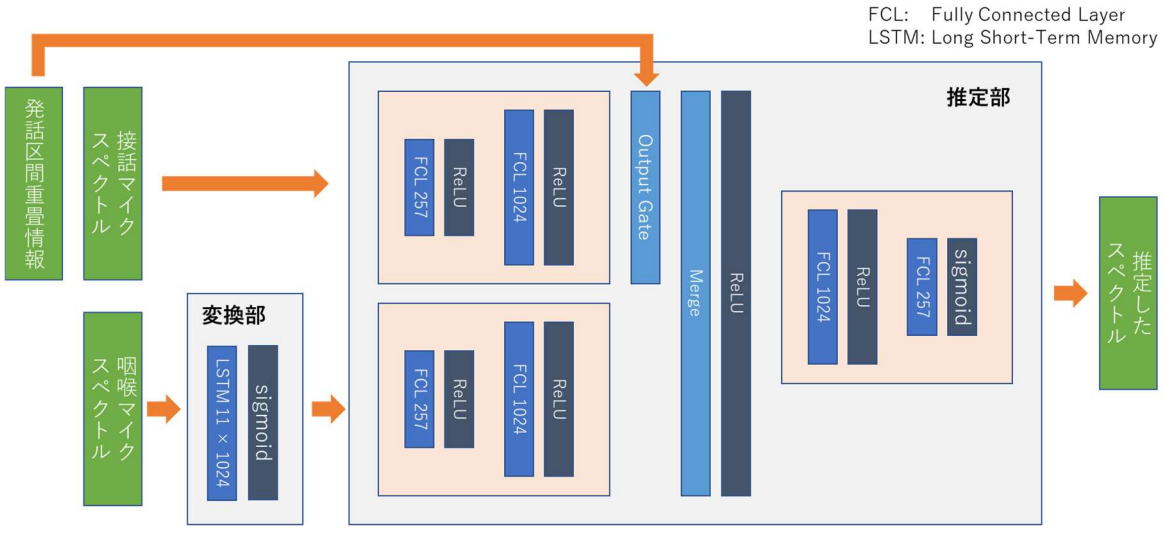

図 1　DNN のネットワーク構成

---

* A Study on Speech Recognition Using Multiple Wearable Microphones for Multi-party Conversations, by LIN, Shengke and TSUNAKAWA, Takashi and NISHIDA, Masafumi and NISHIMURA, Masafumi (Shizuoka University).

3　評価実験

3.1 実験条件

グループワークの状況を模した大語彙音声認識タスクを用いて認識性能の比較を行う．テストデータは男性話者 10 名がそれぞれ異なる資料を読み，ほかの話者に対してその内容を説明させて収集した自由発話(22 分)である．　一方 DNN の学習には男性話者 8 名による音素バランス文 1056 文(77 分)を用いる．音声は防音室にてサンプリング周波数 16kHz で録音し，後処理で発話を重畳させた．なお別途、学生 4 人にテーマを与えてディスカッションを行った多人数会話を分析したところ，重畳区間は全体の 36.1%程度，重畳時の接話マイクの SNR は 15.1dB，咽喉マイクの SNR は 26.4dB であったため、この数値に合わせて重畳を調整している．

多人数会話における発話区間，発話重畳区間の推定方法の一つに先行研究[2]で提案された手法があるが，咽喉マイクを用いて，この手法を検証したところ 98.2%の精度で重畳区間の検出を行えた．本論文では特に DNN による多人数会話の認識に焦点を当てるため，発話区間の切り出しと発話重畳区間のラベル付けは手作業で行った．接話マイクのゲートの係数は，通常時は 1.0 であるが，発話重畳時は 0.5 とした．短時間フーリエ変換のフレーム長は 512 点，シフト長は 100 点である．

3.2 スペクトル分析

クリーン環境と雑音環境における，接話マイクと咽喉マイク，提案手法のスペクトルを図 2 に示す．

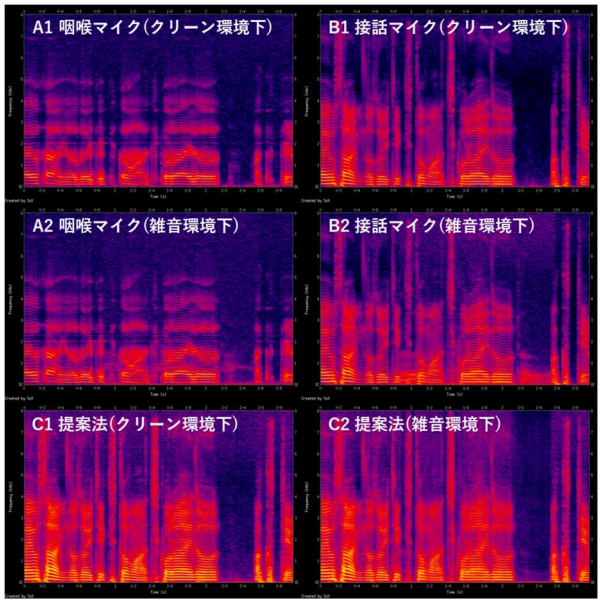

図 2　音声スペクトルの比較

スペクトルを見ると B2 で重畳された発話が, C2 で除去されているのが分かる．また B1 と C1 はほぼ同じであり，DNN によりスペクトルが大きく変質してしまうこともなかった．

3.3 認識実験結果

表 1 に大語彙音声認識タスクにおける認識性能を文字誤り率(CER)で示す．クリーン環境において提案法は接話マイクとほぼ同等の認識性能を示し，雑音環境では接話マイクに比べて約 12%の性能改善を示した．特性の異なるマイクを複数組み合わせることにより，変換に伴う性能低下もなく，発話重畳時の認識性能の改善を達成できた．

表 1 大語彙音声認識の結果

	クリーン環境下 CER	雑音環境下 CER
接話マイク	36.1%	45.3%
咽喉マイク	76.6%	78.5%
提案法	36.0%	40.0%

4　おわりに

接話マイクと咽喉マイクで同時収録した音声を入力とし、DNN でクリーン部分の推定を行うことにより，発話重畳時における認識性能の向上が確認できた.今後は発話重畳の検出も含めたより実環境に近い環境での検証を行う．

謝辞

本研究の一部は科研費（16H01817）の助成を受けた．

参考文献

[1] 林升柯，西田昌史，西村雅史, “多元的音情報に基づく頑健な音声認識に関する研究”，音講論(春), pp 159-160, 2016.

[2] 大高祥裕，綱川隆司，西田昌史，西村雅史, “咽喉マイクとピンマイクの同時集音に基づく多人数会話における発話区間推定に関する研究”，信学技報, vol. 116, no. 279, SP2016-43, pp. 15-20, 2016.