G-1

# 骨導音声の認識率向上の検討
## -深層学習を用いたスペクトログラム補正-
## Improvement of Recognition Rate of Bone-conducted Speech
## -Spectrogram Correction using Deep Learning-

○藤岡紘展[1]，関弘翔[2]，細野裕行[2]
*Hironobu Fujioka[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: Demand for speech recognition using bone-conduction microphone with excellent environmental noise reduction is expected in the future. However, the accuracy of bone-conducted speech recognition using a speech recognition model for air-conducted speech is low. In this report, we study the improvement of recognition accuracy by learning the correction of bone-conducted speech and air-conducted speech using deep learning, and evaluate our method through experiments.

## １．まえがき

近年，交通・工事騒音などの増加に伴い，高騒音環境下かつ防護具使用時における音声伝送の手段として，環境雑音の低減性に優れた骨導マイクロホンの利用が提案されてきた[1]．また，音声認識デバイスの普及から，骨導マイクロホンを利用した音声認識の需要が今後見込まれる．しかし骨導マイクロホンは，気導マイクロホンに比べて収音できる帯域が狭く特性も異なるため，気導マイクロホン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音声認識の性能低下が生じる[2]．

そこで本研究では，骨導マイクロホンで録音した音声(骨導音声)を，気導マイクロホンで録音した音声(気導音声)の様な，通常の音響モデルに適した音声に補正することで，骨導マイクロホン使用時における音声認識の性能低下の解消を図る．

本報告では各音声をスペクトログラム化し，深層学習により補正を行うことで，骨導音声の認識率向上を可能とする手法について検討した．

## ２．認識率向上化のための手法

本報告では骨導，気導の各マイクロホンで録音した1単語ずつの音声(各5秒以内)を1秒ずつスペクトログラム化する． 次に，骨導音声スペクトログラムを入力データ，気導音声スペクトログラムを教師データとしてスペクトログラムの補正を学習する．補正後のスペクトログラムから再構成した音声を，Google 社が提供する音声認識エンジン「Cloud Speech to Text」[3]で認識することで音声認識の精度を比較する．

### ２．１．各音声仕様とスペクトログラムの条件

各音声仕様及びスペクトログラムの条件を以下に示す．また，録音に使用した機材を Table 1 に示す．

[気導，骨導音声]
・サンプリング周波数 ：30kHz
・ファイル形式 ：.FLAC
[スペクトログラム条件]
・FFT 窓 ：0.02 秒
・窓関数 ： ハミング窓

**Table 1.** Used equipment

| | Company name | Product name |
|---|---|---|
| Microphone | Mic W | Air-conduction microphone (I456 Microphone) [4] |
| | Temco Japan | Bone-conduction microphone (EN21N-Tip) [5] |
| IC recorder | ZOOM | H6 Handy Recoder |

### ２．２．学習モデル

骨導音声スペクトログラムはノイズの乗った気導音声スペクトログラムと仮定し，本報告では骨導音声スペクトログラムを気導音声スペクトログラムに補正する手法として画像ノイズ除去のための CNN を参考に検討する．画像ノイズ除去 CNN の参考モデルとして，Fig. 1 に示すモデル[6]を用いた．このモデルは，特徴を集約してサイズを小さくすることで空間的な拡張を得る Pooling を用いないため，音声処理において重要な時間軸方向の詳細なコンテキストが保持できる．

新たに構築する提案モデルを Fig. 2 に示す．提案モデルは，参考モデルと同様の畳み込み(Conv)層，バッチ正規化(BN)層，活性化関数(ReLU 層)の繰り返しに加え，特徴マップのサイズを小さくせずに空間的な拡張が得られる Dilated Convolution[7]と，多層のネットワークの効率的かつ効果的な学習に有効な Residual Block[8]の Shortcut Connection 構造を取り入れた．

１：日大理工・院(前)・情報　２：日大理工・教員・情報

２．３．学習条件

本実験で発音する単語データベースは東北大学 –松下 単語音声データベース(TMW)[9]を参考に，全ての音素環境が出る単語 212 語，鉄道駅名・線名 2788 語で構成した．そのうち 350 単語をテスト用データ，2650 単語を学習用データとした．発音話者は 20 代男性 1 人のみで，各単語 5 秒以内で録音した．

また，学習の損失関数には平方根平均二乗誤差(RMSE : Root Mean Squared Error)を用いる．RMSE は以下の式(1)で算出する．$N$ はデータ数，$x$ は学習用データ，$t$ はテスト用データを示す．

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(x_i - t_i)^2}{N}} \quad (1)$$

３．結果

本報告における音声の認識結果に用いる指標として単語正解率(PC : Perfect Correct)[10]を用いる．単語正解率の算出式を式(2)に示す．*all* は全単語数，*cor* は正しく認識した単語数を示す．

$$PC = cor/all * 100 \quad (2)$$

参考モデルと提案モデルでの学習時 RMSE と，気導音声，骨導音声及び再構成した音声の認識結果を Table 2 に，気導音声，骨導音声及び再構成した音声のスペクトログラムを Fig. 3 に示す．Table 2, Fig. 3 より，提案モデルのほうが参考モデルに比べ RMSE が高く，スペクトログラムはほぼ変わりはないが，認識結果が高いことがわかる．以上のことから，骨導音声の認識率向上の手法として，Dilated Convolution と Shortcut Connection を導入した提案モデルが有効であると考えられる．

４．まとめ

本報告では骨導音声の認識率の低さに着目し，深層学習を用いたスペクトログラム補正による骨導音声の認識率向上手法について検討した．実験の結果，Dilated Convolution と Shortcut Connection を用いることで骨導マイクロホンと通常の音響モデルとのミスマッチを減らすことができ，認識率を向上することができた．

今後は複数人の音声での認識率向上を目指し，モデルの改良とデータセットの改良を検討する．

５．参考文献

[1] 前田秀彦，他：「骨固定型ピックアップから導出した直接骨導音の音響特性」，音声医学言語学会, 2016

[2] 林升柯，他：「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」，情報処理学会, 2016

[3] Cloud Speech-to-Text ホームページ, https://cloud.google.com/speech-to-text/?hl=ja (2019-09)

[4] Mic W ホームページ, http://www.mic-w.com/product.php?id=24 (2019-09)

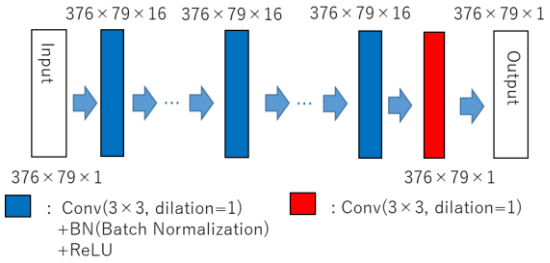


**Figure 1.** Reference model

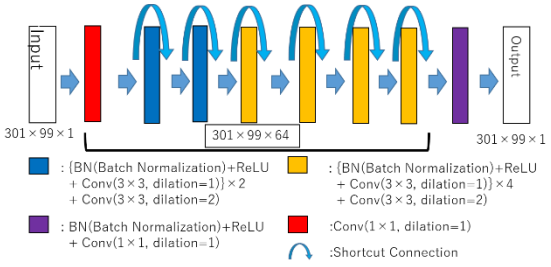


**Figure 2.** Proposed model

**Table 2.** Recognition Result

| | Air-conducted speech | Bone-conducted speech | Reference model | Proposed model |
|---|---|---|---|---|
| Validation RMSE | | | 6.9891 | 6.9921 |
| Test Perfect Correct[%] | 95 | 40 | 63 | 67 |

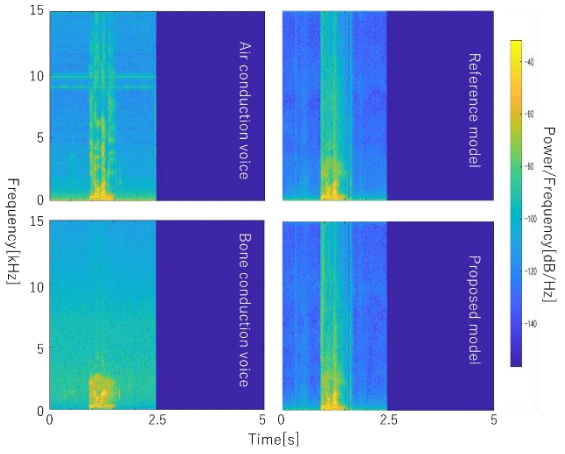


**Figure 3.** Spectrogram Comparison(Speak : “irogami”)

[5] TEMCO ホームページ, http://www.temco-j.co.jp /pr_hg42tbt/ (2019-09)

[6] K. Zhang, et al, “Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising,” IEEE Transactions on Image Processing. Vol. 26, Issue 7, pp. 3142–3155, 2017

[7] F. Yu, et al., “Multi-Scale Context Aggregation by Dilated Convolutions,” ICLR 2016, 2016

[8] K. He, et al., “Deep Residual Learning for Image Recognition,” IEEE Conf. on CVPR 2016, 2016

[9] 東北大-松下 単語音声データベース (TMW), http://research.nii.ac.jp/src/TMW.html (2019-09)

[10] 篠田浩一：「音声認識」，講談社(2017)