第 6 回 東北地区音響学研究会 (2023 年 11 月 25 日) 資料番号 6-5

大規模事前学習モデルに基づく音声認識による日本語単語了解度推定法の検討 *

◎服部真稀, 近藤和弘 (山形大院・理工学研)

1 はじめに

雑音や残響などによる劣化音声に対する人間の聞き取り能力は了解度 (intelligibility) で測定されるが, その主観評価試験は非常にコストがかかる. その為, 人間の心理値である主観評価値を音声の物理量から推定する客観評価手法がこれまで研究されてきた. 音声了解度推定 (Speech intelligibility prediction; SIP) タスクの実現において, 劣化音声の評価の際にリファレンス信号としてクリーン音声を必要とする侵入型 (intrusive) の手法に比べて, 劣化音声の入力のみで了解度を推定する非侵入型 (non-intrusive) の手法は実用的であるが推定難易度は高い. また, 主観的な了解度データは集まりにくく, モデルの構築のコストが大きいという課題もある. 我々はこれまでに音声認識 (Automatic Speech Recognition; ASR) の SIP モデルとしての利用を検討してきた. 音声認識は音声の入力だけで機能し, かつ音声とその転記のみでモデルを構築できるので従来の SIP の問題点の解消できる. しかし, この SIP の推定能力は音声認識の認識精度や音声表現力に依存する. 例えば Hybrid 型 ASR による先行研究 [1] では, 主観評価に対して比較的高い相関係数が見られたものの誤差は大きくなった. また, 話者や雑音条件ごとに適応学習が必要であり, 単一のモデルの汎化性能はあまり無いとされる. そこで本研究ではより高精度で汎用的な音声認識として事前学習モデルに基づく最新の音声認識モデルを利用した了解度推定を試みた. 主観評価, 先行研究との比較から将来の汎用的な了解度推定モデルとしての検討を行う.

2 大規模事前学習モデル

DNN の以降, 様々な機械学習手法が登場したが, 近年の急速な進展と応用を実現しているのが Transformer [2] である. 音声分野においてもその傾向は見られる. 特に自己教師あり学習で行う Transformer の音声表現学習によって, 汎用的な音声処理タスクに有効なエンコーダモデルを実現できるようになった. このエンコーダは事前学習モデルと呼ばれ, 自己教師あり学習 (Self Supervised Learning; SSL) は大量の音声データで学習させることが重要であることから, 大規模事前学習モデルともいう. 音声認識においては事前学習モデルに小数のラベル付き音声でファインチューニングするだけで高性能なモデルを実現できることが利点の 1 つであり, 例えば音声サンプルの少ない言語や方言の音声認識の実装で有効である. SIP モデルの実装に際しても, 集まりにくいとされる了解度試験のデータセットを有効活用でき, SIP タスクにおけるデータ不足の解消が期待できる. 本研究では wav2vec 2.0 [3] と HuBERT [4] の 2 つのアーキテクチャを使って実験をした. 以下に各モデルの大まかな事前学習の仕組みと推論時の流れを紹介する.

2.1 wav2vec 2.0

2 つのアーキテクチャで大まかな流れは共通であり, CNN モジュールで生の音声波形から音声の潜在表現を学習し, Transformer モジュールで潜在表現をコンテクスト表現として学習する. 違いとしては量子化モジュールと全体のネットワークの End-to-End 学習にある. 図 1 に示すように, wav2vec 2.0 では音声の潜在表現は量子化モジュールを通じて符号化ベクトルに変換される. 全体では各フレームの類似度を測る対照学習が行われている. これによって実現された wav2vec 2.0 の事前学習モデルは前述の通り, スクラッチから学習した従来の音声認識モデルに比べて少数のデータセットで高い性能が実現される. また, XLSR-53 や XLS-R [5] のような多言語での学習の有効性が示されており, マルチリンガルにおいて共通の事前学習モデルを利用することが出来るとされる. 一方で量子化モジュールの安定性や妥当性については議論されており, 実際に筆者も学習や出力が安定しないことを確認して例がある.

2.2 HuBERT

HuBERT では wav2vec 2.0 における量子化モジュールは予め MFCC などの特徴量で学習された k-means で構成されている. 符号化ベクトルに対して, k-means の出力は擬似ラベルと呼ばれる. HuBERT の全体の学習は 2 段階に分かれており, 1 段階は初期学習として擬似ラベルと潜在表現によるマスク予測学習によって Transformer のネットワークを学習する. 2 段階以降の反復学習では, 学習済み Transformer の出力を k-means に通すことで得られる疑似ラベル系列を教

* Estimation of Japanese word intelligibility by speech recognition based on a large-scale pre-training model.
by HATTORI, Masaki, KONDO, Kazuhiro (Yamagata University)

師ラベルとして同様の学習を複数回行い, 疑似ラベルを更新する. HuBERT は k-means の導入によって学習目標が明確になり, 純粋な識別基準によって学習を行うことから, 学習の安定性や推論の速さで利点がある. wav2vec 2.0 に対しては同等以上の音声認識性能があるとされている.

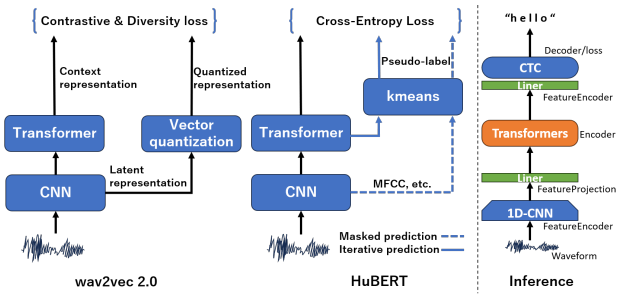

Fig. 1 wav2vec 2.0 and HuBERT based self-supervised learning and inference

3 主観評価試験 JDRT

JDRT は単語対の先頭子音の聞き分け能力の診断によって了解度を測定する主観評価試験方法である [6]. 試験は先頭音素のみが異なる単語対で二者択一の試聴試験を繰り返し, その正解率を了解度とする. 単語は子音属性によって 6 つに分類され, 各特徴につき 10 ペア, 合計 120 単語から構成される. 以下は 6 つの子音属性とそれらの特徴である.

- Voicing: 有声音と無声音の分類 (ザイ：サイ)
- Nasality: 鼻音と口音の分類 (マン：バン)
- Sustention: 継続性の有無 (ハシ：カシ)
- Sibilation: 波形の不規則性の分類 (ジャム：ガム)
- Graveness: 抑音と鋭音の分類 (ワク：ラク)
- Compactness: エネルギーの集中 (ヤク：ワク)

正解率 S は式 (1) によって算出し, R, W, T はそれぞれ正答数, 誤答数, 総試験回数である. S は 100% から-100%の値をとり, ランダムな回答は理想的には 0%に収束する.

$$S = \frac{R - W}{T} \times 100[\%] \tag{1}$$

Kaga2006 は JDRT の主観評価試験データセットであり, JDRT 単語が収録された音声と了解度のスコアから成る. 本研究ではモデル構築での学習データと評価の際に, Kaga2006 を使用した. 音声の条件を Table 1 に示す.

Table 1 JDRT speech dataset (Kaga2006)

Items	Condition
Speaker	4 males & 4 females
Word	60 pairs, total 120 words
Audio	1 ch/16 bit/ 16kHz
Noise	Clean Speech, Speech-Shaped Noise White Noise, Bubble Noise
SNR	10, 0, -10, -15 [dB]
Total	12480

4 実験条件

実験の流れである音声認識モデルの作成, モデルでの JDRT 模擬, モデルの評価を順に解説する.

4.1 モデル

wav2vec 2.0 または HuBERT の各事前学習モデルを JDRT 音声でファインチューニングすることで, 言わば”JDRT 用の音声認識モデル”を作成している. 2 つのモデルアーキテクチャに対してそれぞれ 2 つの学習条件で学習する為, 合計 4 つのモデルを作成した. 利用した事前学習モデルは wav2vec 2.0 の日本語ファインチューニングモデル [7] と HuBERT の日本語事前学習モデル [8] である.wav2vec 2.0 には [8] のように日本語での自己教師あり学習による事前学習モデルは現時点では存在しない為, 英語の自己教師あり学習による事前学習モデルを日本語でファインチューニングしたものを本研究で使うことにする. 次に学習条件を Table 2 に示す. 学習データは音素のラベルが付与された JDRT 音声であり, CTC デコーダの出力音素列の正解を目標に学習する. 学習に使用した JDRT のクリーン音声 (clean*) には 7 パターンの話速変換による Data Augmentation が適用されている. Sox コマンドの speed オプションによって話速を {0.85, 0.90, 0.95, 1.00, 1.05, 1.10, 1.15} 倍速で変化させた. Noisy のデータセットでは Kaga2006 の条件に加えてピンクノイズとブラウンノイズを含めた 5 種類の雑音で加算した. 当初 JDRT 音声のみでの学習を試みたが, データの偏りなどから学習や ASR の出力が安定せず, 親密度別単語了解度試験用音声データセット FW07 [9] を混ぜて学習させることで安定性を得た. 学習データとテストデータ (Kaga2006) はデータ自体の重複はないが, 雑音種, SNR, 発話内容においてオーバーラップしていることになる.

4.2 JDRT 模擬

作成した音声認識モデルに Kaga2006 の音声を入力し JDRT を模擬する. 本来の人間の回答としては単

Table 2　Training Condition

Models		Epochs	Train dataset
clean	wav2vec 2.0	10	・clean* JDRT(6720)
	HuBERT	20	・clean FW07(6400)
noisy	wav2vec 2.0	15	・clean*/noisy JDRT(25920)
	HuBERT	20	・clean/noisy FW07(134400)

語を選択することになるが, 音声認識の回答としては各音素の確率, またはその最大化による最終的なトークン列である. 実験全体を通じて音声認識から明確なトークン列が得られたことから, そのトークン列の冒頭音素の正誤を判定した. 各音声に対する正誤を集計し, 雑音種, SNR, 子音特徴の条件の組み合わせにより C 式 1 に基づいて 90 サンプルの CRR を算出する.

4.3　評価

モデルの出力から算出される CRR と Kaga2006 で同様の条件で算出される CRR の相関・誤差を取ることで提案モデルの SIP としての評価を行う. 相関係数としてピアソンの積率相関 (Linear Correlation Coefficient; LCC) と, スピアマンの順位相関 (Rank Correlation Coefficient; RCC) の 2 つ, 誤差として二乗平均平方誤差 (Root Mean Squared Error; RMSE) を用いた. また, 現時点でのモデルの汎用的な ASR の性能評価として CSJ コーパス [10] の eval 評価セットに対する音素誤り率 (Phoneme error rate; PER) を測定した. なお, 評価セットは単語では無く文章認識タスクである.

5　結果と考察

Table 3 に 4 つのモデルから推定した了解度と主観評価との相関, 誤差を示す. また, 先行研究 [1] である GMM-HMM の結果との比較をする. wav2vec 2.0 と HuBERT で共に clean に対して noisy のスコアが上回った. GMM-HMM で行われた雑音のマルチコンディション学習は本研究の noisy と同等の学習状況と考えられるが, wav2vec は同等の精度が得られ, HuBERT に関しては大きく上回る結果となった. これらの結果は良い推定モデルを構築出来たと評価出来る一方で, 人間の雑音に対する反応を表現したのではなく, 単なる機械学習のロバストネスな結果である可能性もある. 今後は雑音に対するクロスバリデーションや未知雑音, 非定常雑音に対する反応などを評価する必要がある. 先行研究での適応学習時に見られた雑音学習による原音声認識率の低下や, 顕著な話者依存性は wav2vec 2.0 や HuBERT では無く, 大規模事前学習モデルの特徴表現の汎用性と効果的なファインチューニングの結果によるものだと考えられる.

Table 4 は音素誤り率の結果である. CSJ のスコアは 3 つの評価セット (eval1, 2, 3) の平均値である. JDRT は認識単語と雑音のドメイン適応が取れており, 更に学習によって向上が見られる. CSJ の評価セットでは高い精度とは言えず, 雑音で調整されたモデルによって値は増加した. しかし, 単語だけを学習したモデルである点, 言語モデルを使っていない点を考慮する必要がある. 今後の検討としては音声認識と了解度推定を同時に推論するようなマルチタスクモデルの実装によって改善が見込めると考えている.

Table 3　Correlation and error between subjective and objective evaluation

Models		LCC	RCC	RMSE
wav2vec 2.0	claen	0.81	0.89	36.15
	noisy	0.95	0.95	15.42
HuBERT	claen	0.86	0.93	26.47
	noisy	0.97	0.95	8.41
GMM-HMM	multi	0.97	0.93	14.14

Table 4　PER for each evaluation dataset

Models		CSJ eval	JDRT Kaga2006
wav2vec 2.0	claen	0.42	0.39
	noisy	0.49	0.14
HuBERT	claen	0.39	0.29
	noisy	0.43	0.07

6　おわりに

本研究では音声認識を用いた音声了解度の客観評価指標の検討として, wav2vec 2.0 や HuBERT などの自己教師あり学習に基づく大規模事前学習モデル用いて音素認識モデルを構築し, 単語や試験設定を限定した了解度推定を行った. 構築した音声認識モデルの出力と人間の了解度において高い相関が得られ, 加算雑音でのファインチューニングによってその精度は向上した. 事前学習モデルと少数のデータセットで了解度推定モデルを構築出来ることは利点である. しかし, 機械学習では学習データに依存することは避けられない為, 学習や検証において今後の更なる調査が必要である. また, 今後の検討としては音声認識モデルの潜在表現と利用することである. 音声認識による音声表現と了解度に関連を見つけることが出来ると, 単語や試験など学習状況に依存しない理想的な了解度推定の実現につながると考えている.

参考文献

[1] T. Yusuke *et al.* Estimation of speech intelligibility using speech recognition systems. *IEICE TRANSACTIONS on Information and Systems*, Vol. 93, No. 12, pp. 3368–3376, 2010.

[2] A. Vaswani *et al.* Attention Is All You Need. In *NIPS2017*, 2017.

[3] Alexei Baevski *et al.* wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[4] Wei-Ning Hsu *et al.* Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[5] Q. Xu *et al.* Simple and effective zero-shot cross-lingual phoneme recognition, 2021.

[6] K. Kondo *et al.* On a two-to-one selection based japanese speech intelligibility test. *Acoust. Sci. & Tech.*, Vol. 63, No. 4, pp. 196–204, 2007.

[7] Jonatas Grosman. Fine-tuned XLSR-53 large model for speech recognition in Japanese. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-japanese`, 2021.

[8] rinna Co., Ltd. japanese-hubert-base. `https://huggingface.co/rinna/japanese-hubert-base`, 4 2023.

[9] S. Amano *et al.* NTT-Tohoku University familiarity-controlled word lists 2007. *Speech Resources Consortium, National Institute of Informatics*, 12 2007.

[10] K. Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.