発話障害者の自発音声を対象とした非参照型明瞭度予測 *

○ HALPERN Bence Mark, 戸田 智基 (名古屋大学)

1 はじめに

自然発話（Spontaneous speech）の評価は、朗読発話と比較して日常のコミュニケーション能力をより正確に反映するため、発話障害の臨床評価において極めて重要である。しかしながら、既存の客観評価手法の多くは、病的な自然発話に対して取得が困難な参照テキスト（reference transcripts）に依存している。本研究では、自動音声認識（ASR）を用いた参照を用いない明瞭性評価手法を提案する。本手法では、ASRを用いて擬似的な正解テキスト（Pseudo-reference）と認識仮説（Hypothesis）の両方を推定し、人手によるアノテーションなしに明瞭性を算出する。口腔癌患者の音声コーパスを用いた実験により、主観評価値と比較して提案手法の有効性を確認した。

2 背景

発話明瞭性の評価は、口腔癌、舌切除、あるいは構音障害などに起因する病的音声のリハビリテーションおよび診断において基本的なプロセスである。臨床家（言語聴覚士など）による主観評価はゴールドスタンダードであるが、多大な時間を要し、評価者間でのばらつきが生じるという問題がある。そのため、信頼性の高い客観的尺度の開発が研究の主要な焦点となっている。

ASR から導出される単語誤り率（WER）[1] や文字誤り率（CER）などの標準的な客観尺度は広く使用されている。しかし、これらの指標には重大なボトルネックがある。それは参照テキスト（reference）を必要とすることである。自然発話の文脈において、正確な参照を取得することは困難である。患者の発話には、非流暢性、文法的な破綻、深刻な構音誤りが含まれることが多く、人間のアノテーターであっても「正解」の書き起こしを定義することは曖昧になりがちである。

近年の End-to-End 機械学習モデルは、音響特徴量から直接スコアを予測することで参照を用いない解決策を提供しているが、説明可能性（Explainability）に欠ける場合が多い。これらの「ブラックボックス」的なアプローチは、患者の発話がなぜ不明瞭なのか（例：特定の音韻誤りなど）という解釈可能な洞察を臨床家に提供することができない。

これらの制限に対処するため、我々は参照を用いない、かつ解釈可能な明瞭性予測フレームワークを提案する。音響特徴量のみに依存する（信号ベースの）非侵襲的な手法とは異なり、本手法は最新の ASR の言語処理能力を活用する。我々は、ASR モデルの「意図した」予測（擬似正解）と「知覚した」認識結果との間の乖離が、人間の明瞭性評価の解釈可能な代替指標となりうると仮説を立てた。

3 提案手法

本手法の基本的なアイデアは、2 つの異なる ASR モデルの利用にある。第 1 のモデルは「知覚された音声」の書き起こしを目指し、第 2 のモデルは「意図されたメッセージ」の書き起こしを目指し、これを参照として機能させる。

これを計算機的に実装するため、異なるアーキテクチャ上のバイアスを持つ 2 つの ASR モデルを採用する。1 つ目のモデル A（M_{ref}）は、強力な言語モデリング能力または歪みに対する頑健性を持ち、意図されたメッセージ（擬似正解：Pseudo-Reference）の復元を担う。2 つ目のモデル B（M_{hyp}）は、音響的な詳細に対してより敏感であり、知覚された音声（仮説：Hypothesis）の書き起こしを担う。

形式的には、音響特徴量の入力系列を X とする。2 つの ASR モデルを M_{ref}、M_{hyp} と定義する。これらは以下の通り 2 つの異なる書き起こしを生成する。

$$T_{ref} = M_{ref}(x) \quad T_{hyp} = M_{hyp}(X) \tag{1}$$

明瞭性スコア S は、これら 2 つの推定値間の乖離を定量化することで導出される。発話が明瞭であれば、両モデルは類似した書き起こしに収束するはずである。発話に障害がある場合、音響に敏感なモデル（M_{hyp}）は誤りをそのまま書き起こす可能性が高い一方、文脈に頑健なモデル（M_{ref}）は誤りを補正しようとするため、両者の乖離は大きくなる。

スコアは以下のように計算される。

$$S = \mathrm{Sim}(T_{ref}, T_{hyp})$$

ここで、Sim は類似度または距離尺度を表す。本研究では、T_{ref} と T_{hyp} の間の音素誤り率（PER）および単語誤り率（WER）を計算する。これらの発話レベルのスコアを平均化し、録音レベルのスコアを得る。

---

* Reference-Free Intelligibility Prediction for Spontaneous Speech of People with Speech Disorders using ASR-based Text Estimation by HALPERN, Bence Mark; TODA, Tomoki(Nagoya University)

Table 1 Pearson correlation of recording-level scores with subjective ratings of human listeners compared. **Boldface** indicates best result.

Method	r_{PER}	r_{WER}
Reference-based（参照あり）		
wav2vec	0.63	0.39
DNN-HMM	0.80	0.69
Reference-free（参照なし・提案法）		
wav2vec + DNN-HMM	**0.82**	0.71

4 実験的評価

4.1 データセット

提案手法の評価には YouTube Oral Cancer Corpus [2] を用いた。これは、何らかの口腔癌手術を受けたとされる 21 名の話者による、13 の自然発話録音（アメリカ英語）から構成されている。主観的な正解値（Ground Truth）を収集するために、各録音から複数の発話が抽出された。これらの発話に対し、100 名の非専門家（アメリカ英語話者）と 1 名の専門家が 5 段階評価を行った。各録音の全体的な明瞭性スコアを得るために、個々の発話評価を集約した。先行研究において非専門家と専門家の評価に高い相関が示されているため、本研究では統計的な信頼性が高い非専門家による集約スコアを正解値として使用した。

4.2 実験設定

wav2vec ベース [3] の ASR モデルと、YouTube Oral Cancer Corpus [2] 上で構築された n-gram ベースの最先端 Kaldi DNN-HMM ハイブリッドシステムを使用した。ベースラインモデルには、書き起こしが存在するすべての録音でテストされた先行研究のものを使用した。

評価においては、まずモデル予測と正解値（Ground Truth）を比較してベースラインを確立し、その後、参照を用いない予測同士を比較した。主観評価のスキームに合わせるため、客観スコアの算出は発話レベルで行った。各発話について個別に PER と WER を計算し、それらを集約して録音レベルのスコアを算出した。正書法の書き起こしのみが得られているため、*phonemizer* ライブラリ [4] を使用して書記素-音素変換を行った。

5 結果

表 1 に客観尺度と人間の主観評価との間のピアソン相関係数（r）を示す。提案手法は 0.82 の音素相関を達成し、それぞれ単独の reference-based のシステムを上回った。これは、本システムが手動の書き起こしを必要とせず、音響信号と ASR 言語モデルの相互作用のみから明瞭性の特徴を捉えることに成功したことを示唆している。

6 おわりに

本稿では、発話障害者の自然発話に対する参照を用いない明瞭性予測手法を提案した。ASR を用いて参照と仮説の両方を推定することで、手動による書き起こしの必要性を排除した。口腔癌データセットを用いた実験により、提案手法が人間の知覚と高い相関を持つことが示された。今後の課題として、他言語への拡張およびリアルタイムの臨床フィードバックアプリケーションへの統合が挙げられる。

謝辞 本研究の一部は，NWO（プロジェクト番号 019.232SG.011），及び，NEDO の委託事業である JPNP25006 の支援を受けたものである．

参考文献

[1] A. Maier et al., “Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–7, 2009.

[2] B. M. Halpern, S. Feng et al., “Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners,” *Speech Communication*, vol. 149, pp. 84–97, 2023.

[3] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] M. Bernard and H. Titeux, “Phonemizer: Text to phones transcription for multiple languages in python,” *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.