クラウド汎用音声認識APIを用いた日本語単語了解度の推定*

◎服部真稀，近藤和弘（山形大）

1　研究背景・目的

通信技術やデバイスの向上によって，様々な環境下での通話や音声サービスが可能となった．その中で雑音や残響の混入によってその快適さが損なわれる可能性があり，音声の品質保証が必要となる．本研究で扱う了解度(Intelligibility)は「音声情報の伝わりやすさ」を表す音声品質指標の１つである．

音声了解度を求める方法は大別すると，主観評価法と客観評価法の 2 種類である．試聴試験によって測られる主観評価は最も信頼できるのだが，多数の被験者を使った長時間の試験を実施する必要がある為，コストや安定性が問題となる．一方で了解度を高精度で推定する客観評価の実現はその問題を解消することができる．

我々は主観評価手法である日本語版 DRT(Diagnostic Rhyme Test)[1]とその客観評価方法について研究してきた．劣化音声のみから推定するノンレファレンス式の客観評価方法においては様々な推定器が検討されたが，音声認識を推定器として用いる手法はその１つである．手法[2]では GMM-HMM (Gaussian Mixture Model – Hidden Markov Model)を手法[3]では DNN-HMM(Deep Neural Network)を音声認識エンジンとして了解度を推定したが，主観評価の傾向を完全に再現出来てはいない．現在，これら Hybrid 型音声認識の適応学習や，End to End 音声認識の採用によって更なる精度向上を狙っている．

現在, Google や Microsoft, Amazon などの企業は大規模な音声・言語データによって構築された高精度な音声認識 API を提供している．我々は将来的に単語制約を無くし，リアルタイム性を備えた客観評価システムを目指しており，これら汎用 API と Hybrid 型との知識蒸留を実現方法として検討している．本研究ではその為の基礎研究として，汎用音声認識 API による日本語版 DRT の模擬を行う．

2　主観評価

2.1 日本語版 DRT

DRT とは語頭 1 音素のみ異なる単語対の聞き分けを行う主観評価試験である．被験者は単語対から 1 単語のみを聴取し，二者択一式で回答する．各単語の語頭子音には 6 つの属性で区別され，以下が 6 つの属性とそれらの特徴である．

Voicing: 有性音と無声音の分類
Nasality: 鼻音と口音の分類
Sustention:継続性のある音と無い音の分類
Sibilation: 波形の不規則性に関する分類
Graveness: 抑音と鋭音の分類
Compactness: スペクトルのエネルギーが１つのフォルマントに集中するか否かの分類

1 つの属性につき 10 単語対，計 20 単語で試験される．正答率は式(1)を用いて算出される．

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

ここで S は正答率(了解度), R は正答数, W は誤答数, T は全回答数である．

2.2 使用音声

主観評価試験は[4]において行われた．本研究の主観評価，客観評価を通じて全て以下の条件の音声で試験を行っている．

Table 1: 音声

項目	条件
発話者	男女各 4 名，計 8 名
単語数	60 単語対，計 120 単語
録音設定	1ch / 16bit / 16kHz
形式	wav
雑音	無加算，ホワイトノイズ バブルノイズ，疑似雑音
SNR(雑音比)	10, 0, -10, -15 [dB]
総音声数	12480

* Estimation of Japanese word intelligibility using cloud generic speech recognition API, by HATTORI, Masaki and KONDO, Kazuhiro (Yamagata University).

3　客観評価

3.1 Hybrid 型音声認識

音声認識システムで日本語版 DRT を模擬することで了解度を推定している．[2],では HTK と Julius を用いて GMM-HMM のシステムを構築した．不特定話者モデルと話者・雑音適応学習済みモデルのそれぞれで試験が行われている．[3]では Julius ディクテーションキットをベースに DNN-HMM のシステムを構築した．こちらは不特定話者モデルのみ試験となっている．これらの Hybrid 型音声認識は出力が二者択一になるように設定しており，試験する単語ごとに 60 単語対分の言語モデルを切り替えている．

3.2 Microsoft Azure Speech to Text

本研究では汎用 API として Microsoft 社の Azure Speech to Text[5]（以下 Azure STT）を使用した．従量課金制のサービスであり，各種開発環境で使用できるようになっている．機能としては，単語登録機能，音響モデル適応学習機能などがある．各社 API は基本的に言語モデルの設定をすることはできない為，出力を二者択一に限定することはできない．その為，今回は音声認識の自由な出力を許可し，単語登録を適用したものと、単語登録無しでの試験を行った．単語辞書は単語対をひらがなで登録し､試験する音声ごとに切り替える。

API の出力イベントとしては「適合」「不適合」「エラー」の３種が存在し，この中で「適合」のみテキスト（認識結果）が出力される（何らかの言葉を認識した状態）．認識されると N-best 候補の上位 5 つを取得することができ，候補それぞれに対して，テキストと信頼度を得ることができる．実験はクリーン音声(960)に対して行った．音響モデルは不特定話者モデル(independent)である。

開発は Visual Studio 2022 上でフレームワーク.Net core で行った．API のバージョンは Microsoft Cognitive Services Speech(Speech SDK)ver.1.24.2 である．

4　結果

Table 2 は単語登録済みの Azure STT によるクリーン音声 960 個の認識結果であり、N-Best 候補の中で最上位（デフォルト出力）から正答率(了解度)を計算したものである。前述の通り二者択一に限定出来ていない為、出力は漢字、カタカナ、アルファベットなど様々である。この出力の文字通りに正誤判定したのが「未認識」、漢字やカタカナをひらがなに変換したものが「同音」、語頭 1 音での正誤判定が「語頭」である。

出力イベントで不適合となった、未認識の音声は 7 個であった。主観評価では未回答は有り得ないのだが、この未認識を誤答として計算した。5 件の修正加えると 92.5%まで増加する。

Table 3 は主観評価と Hybrid 型音声認識のクリーン音声の認識結果である。今回の Azure STT の結果と比較すると、未知話者モデルとしては良い結果であると言える。話者適応することで更なる精度向上が期待できる。

Table 2:クリーン音声の正解率（単語登録）

正誤条件	正答数	未認識含めず		未認識含む	
		誤答	正答率[%]	誤答	正答率[%]
未変換	892	61	87.19	68	85.80
同音	899	54	88.67	61	87.29
語頭	919	34	92.86	41	91.46
修正(5 件)	924	29	93.91	36	92.50

Table 3: 主観評価, Hybrid 型の結果

Condition	Clean
Subjective	98.01
GMM-HMM	94.38
GMM-HMM (independent)	59.58
DNN-HMM (independent)	76.73

5　まとめ

本研究はクラウド汎用音声認識 API を用いた了解度推定の基礎検討である。音声認識エンジンとして Azure SST を使い、日本語版 DRT を模擬することで了解度推定を行った。結果から未知話者モデルでクリーン音声における高い正答率が得られた。

今後は雑音・残響を加えての実験を行い、さらにモデルへの話者・雑音の適応学習を検討している。

参考文献

[1] K. Kondo, et al. "Two-to-one selection-based Japanese speech intelligibility test", J. of Jap. Acoust. Soc., vol. 63, no.4, p. 196-205, 2007.4.

[2] Y. Takano, K. Kondo, "Estimation of speech intelligibility using speech recognition systems", IEICE Trans. Inf. & Syst., Vol. E93-D, No. 12, Dec. 2010.

[3] M. Hattori, K. Kondo, “Estimation of Japanese word intelligibility using DNN-based speech recognition systems”, ICA2022, ABS-0723.

[4] 加賀類, 他, “客観音声品質評価法 PESQ を用いた日本語了解度の推定方法について”, 平成 17 年度第 5 回情報処理学会東北支部研究会,2006.

[5] Microsoft Azure speech to text, “https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/#features”