3.6.1 先進的音声技術研究室

室長　　河井　恒　ほか16名

グローバルコミュニケーション計画に向けた音声技術の研究開発

■概要

当研究室では、機械を介した音声コミュニケーションの基盤となる音声認識、音声合成、音声対話処理の各技術の研究開発に取り組んでいる。東京2020オリンピック・パラリンピック競技大会までに音声翻訳技術の社会実装を実現することを目指して、実用的な性能を有する多言語の音声認識・音声合成技術の開発を推進した。今年度は特に特定技能人材の日常会話対応への応用を想定してモンゴル語の音声認識及び音声合成システムを試作した。一方、2020年以降の世界を見据えて話者認識技術の研究を行った。

■令和２年度の成果

１．モンゴル語の音声認識・合成技術の開発

モンゴル国は中国とロシアの間に位置し、放牧に適した高燥地と山岳地や砂礫地そしてウランバートルなど都市部からなる人口300万人余りの国家である。相撲の分野では既に多くの出身者を日本に迎え、TV等で日本語を流暢に操る姿は馴染みのものとなったが、今後特定技能制度による人材の増加が見込まれ、モンゴル語での会話をサポートする音声翻訳技術が望まれている。そこで、キリル文字表記のモンゴル語を対象として、モンゴル国において成人男女994名から収集した合計142時間の音声によるコーパスを構築し、ニューラルネットベースのモンゴル語自動音声認識システムを開発した。NICTの公開音声認識評価データセットSPREDS3による評価の結果、令和３年３月時点で一般に利用可能な商用APIサービスと同等の認識精度であることを確認した。並行して、やはりモンゴル国において男性の職業話者による約３時間の音声コーパスを収録し、モンゴル語音声合成システムを試作した。以上の成果を実証実験システムVoiceTra（ボイストラ）に搭載し、一般に公開した。

２．音声認識の性能と人間の文字起し能力の比較

ディープ・ブルーの登場以来、様々な分野で『AIが人間を超えた』との報告が相次いだ。今中長期を終えるにあたり、自動音声認識（ASR）の性能が人間の文字起こし能力を超えたか否かを精密な実験によって検証した。NICTの公開音声認識評価データセットSPREDS2の日本語音声を対象として、ASRと人間の文字起こし能力を速さと正確さの両面で比較した。従来は日本語での検討は行われておらず、また速さは比較されていなかった。

ASRはNICTASRの最新モデルとし、人間は日本語音声書き起こしの熟練者３人と校正者１人の体制とした（図１）。校正者はASRでは原理的に生じ得ないタイプミスを修正する役割で加わった。速さは音声の持続長で正規化した所要時間（RTF）、正確さは単語誤り率をそれぞれ指標とした。人間の文字起こしにはWindows10の日本語IMEを予測入力なしで用い、全てのキーの押下時刻をミリ秒単位で記録することで速さを測定した。

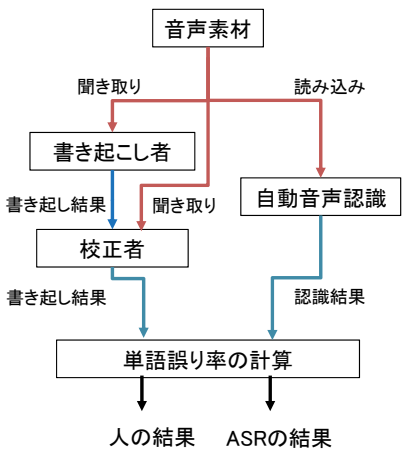

図1　ASRと人間の文字起こし比較実験

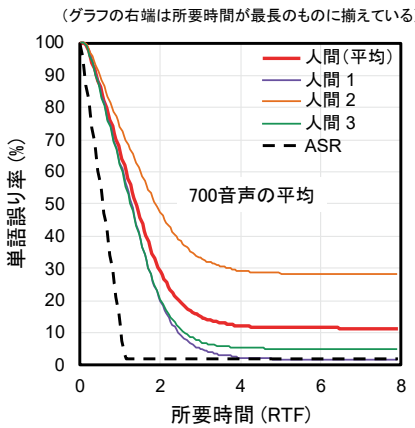

図2　音声再生が1回の場合の結果

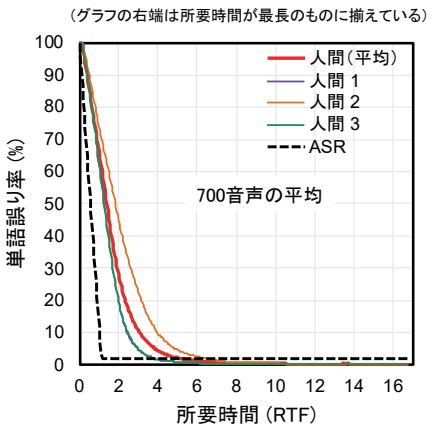

図3　音声再生回数の制限がない場合の結果