合成音声を用いた絵本の読み聞かせにおける聞き取りやすさの評価

青木 茉衣 (福井大学工学部)
小高 知宏・黒岩 丈介 (福井大学大学院工学研究科)
白井 治彦 (福井大学工学部)・諏訪 いずみ (仁愛女子短期大学)

1　はじめに

近年，教育分野にも ICT 技術の導入が進み，合成音声を導入しようとする動きが見られる．音声合成技術は日々向上しており，自然で聞き取りやすい合成音声が作られるようになってきた．しかし，今までこのような合成音声は大人が利用することを想定しており，子どもにとって聞きやすいものは依然少ない．もし子どもにも聞き取りやすい合成音声が手軽に作成できるようになれば，より簡単に子どもたちが読み聞かせを楽しめるようになり，絵本に触れる機会が増えることが期待できる．そこで本研究では，絵本読み聞かせの合成音声システムを作成し，子どもに聞き取りやすい音声か評価を行う．

2　評価手法

本研究では，読み聞かせの対象を 3 歳から 10 歳くらいまでの絵本に親しむ子どもと想定している．対象者にも聞き取りやすい音声は，ゆっくりとした読み上げで発音が明瞭なものである必要があるため，読み上げの速度や間の長さが適当であることが重要だ．そこで間の長さや読み上げの速度を調べ，合成音声が子どもにとっても聞き取りやすいものになっているかを考察する．理想の読み方をナレーターによる読み上げとし，合成した音声の間の長さ及び読み上げの速度の計測を行い，理想との乖離の程度を求める．

評価対象の合成音声は，Google Colabolatory で音声処理ツールセットの ESPnet2[1] を用いて合成した音声である．サンプリング周波数は 24kHz，音響モデルは jsut_conformer_fastspeech2_accent_with_pause，ボコーダーは jsut_multi_band_melgan.v2 を使用した．速度はデフォルト値とした．音声合成システムの手順を示す．環境のセットアップを行い，モデルを設定した後，絵本のテキストを読み込み，ページが変わる箇所でテキストを分割する．次に 1 ページ毎の音声を合成し，それぞれの音声の間に 0.5 秒の無音を挟みながら全ページの音声を結合する．結合した音声合成は wave ファイルとして出力する．

3　実験と結果

実際に合成音声と人による読み上げにおける間の長さ及び読み上げ速度の詳細な計測方法と，結果を以下に示す．

1．間の計測

無音の時間を間とする．絵本作品全体から 5 文選び，各文中の句点と読点それぞれの後ろに存在する無音の時間を計測し，平均値を求める．

表 1: 合成音声と人による読み上げの間の長さの比較

	全77行	間の長さ[s]		比較				間の長さ[s]		比較	
句点	場所	合成音声	人の声	合成-人	合成/人	読点	場所	合成音声	人の声	合成-人	合成/人
1	1行目	0.24	2.693	-2.453	0.089	1	2行目	0.245	1.389	-1.144	0.176
2	16行目	0.26	1.225	-0.965	0.212	2	19行目	0.19	1.315	-1.125	0.144
3	29行目	0.294	2.1	-1.806	0.140	3	38行目	0.143	0.733	-0.590	0.195
4	49行目	0.181	1.484	-1.303	0.122	4	54行目	0.26	1.553	-1.293	0.167
5	67行目	0.216	1.271	-1.055	0.170	5	70行目	0.181	0.743	-0.562	0.244
	平均値	0.2382	1.7546	-1.516	0.136		平均値	0.2038	1.1466	-0.943	0.178

2．読み上げ速度

1 秒間に読む文字数を読み上げ速度とし，絵本作品全体及び 1 文毎の読み上げ速度を計測する．絵本作品全体の速度は作品の 1 文字目から最後の文字まで読み上げる時間を文字数で割ることで求める．また 1 文毎の速度では作品全体から 5 文選び，各文の読み上げ速度を計測して平均値を求める．

表 2: 合成音声と人による読み上げの速度の比較

場所			速度[字/s]		比較	
1文毎	全77行	文字数	合成音声	人の声	合成-人	合成/人
1	1行目	37	8.581	5.089	3.492	1.686
2	18行目	80	8.453	4.696	3.757	1.800
3	40行目	54	8.399	5.603	2.796	1.499
4	52行目	25	9.198	6.112	3.085	1.505
5	70行目	47	7.986	4.818	3.168	1.658
	平均値		8.524	5.264	3.260	1.619
全体	全体	3991	8.129	4.217	3.912	1.928

4　考察とまとめ

結果より，合成音声は人による読み上げより速く，間は短かった．読み聞かせ対象の子どもには速くて聞き取り難いと考えられる．今後は，読み上げ速度や間の長さが人による読み上げに近い合成音声システムを開発する予定だ．当日に詳細な結果も併せて示す．

参考文献

[1] 林知樹,”End-to-End 音声処理の概要と ESPnet2 を用いたその実践”, 日本音響学会誌 vol.76,No.12,pp.720-729 2020