氏　　名	小林 洋介
所属機関	室蘭工業大学
研究題目	「よく聴こえる」拡声システムのための音声合成システムの構築

1. 研究の目的

屋外拡声装置は音声情報を多くの人に伝達するため，生活空間のあらゆる場所に設置されている。しかし，直感的に利用できるが故に最適とは程遠い利用による音割れや相手を意識しないボソボソとした発話により聴こえにくく情報収集が困難になることが多い。一般に聴き取りを改善するため，単純に音響信号にゲインをかける音量調整（いわゆる「ボリュームを上げる」操作）を行うことが多い。しかし，単純に放送地点での音量増加では，拡声フィールドにおける反響・残響が増加し，逆に音声の明瞭性を低下させ，適切ではない。

本研究では，自動処理による聴こえを改善するシステムのプロトタイプの実装と話者性を保持した高音質化の基礎的な統計モデルの検討を行う。

2. 研究の内容（手法、経過、評価など）

2-1. 提案システムのプロトタイプ実装

図1に本研究で実装したシステムのプロトタイプを示す。中段が全体の概要で，入力音声を音声認識(Speech Recognizer)によりテキストに変換し，自然言語処理によるパース処理(Semantic Parser)で必要な情報を抽出し，音声合成部(Speech Synthesizer)で再度合成している。図の下段にある言語処理部は本研究の提案と元となった我々の既往研究の成果である（Y. Kobayashi et al., Proc. GCCE 2017）。本研究においては先行研究では既存の音声合成エンジンを利用していただめ，発話者自身の音声で学習した HMM(hidden Markov model)音声合成を構築した。HMM 音声合成は，学習用音声の音声信号から信号処理で求める音声パラメータとテキスト解析にから求めた言語ラベルにより学習する統計的音声合成モデルである。近年進歩が著しい深層学習による連鎖律を用いた生成モデルの一つである WaveNet と異なり，少ない学習音声でも人間の知覚的に満足できる音質で音声合成が可能な点に特徴がある。

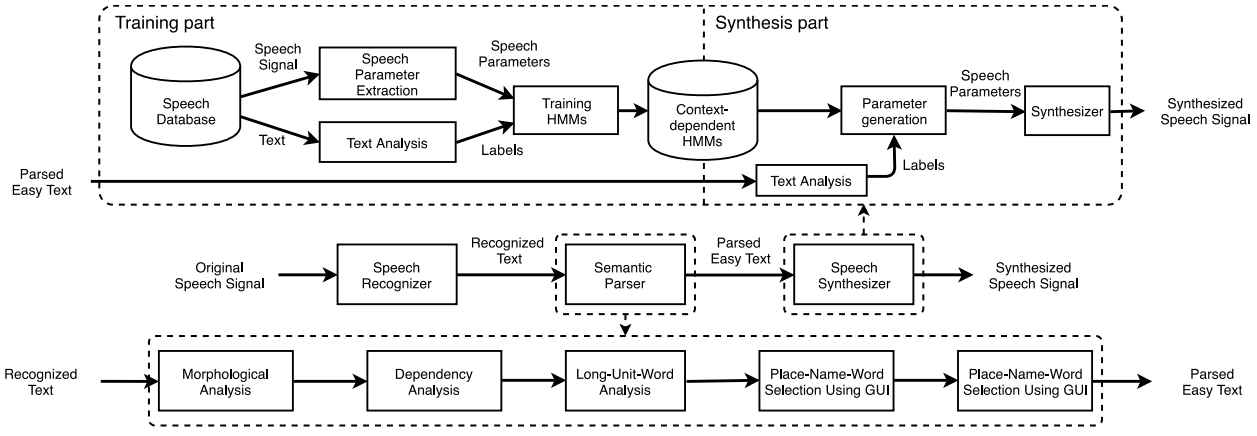

図 1: 提案システムの全体図

2. 研究の内容（続き）（書ききれない場合には、同一形態のページを追加しても結構です）

2-2. Pix2Pix 音声合成を用いた話者変換の実装

HMM 音声合成は統計モデルを利用するが，合成語の音声スペクトログラムが平滑化されている傾向にあり，元の話者性を保持しながらも平均的な声に近くなりやすい。そこで，先述の WaveNet など生成モデルに基づいた統計的音声合成の研究が盛んになっている。生成モデルは，深層学習で作成された音声や画像を種となる符号から生成する技術であり，代表的な手法に敵対的学習である GAN(Generative Adversarial Networks)を用いた画像生成があり，写真のような自然画像も生成できるような統計モデルの応用がある。GAN を利用した画像生成の応用で広く使われる技術に Pix2Pix が 2017 年に提案され，「地図画像と航空写真のようにペアとなる画像を学習することで架空の地図を合成する」，「建物の写真と窓の位置の図面を学習して架空の建物写真を合成する」といった対応を持った画像の変換の提案がされている。

提案システムは元の発話者の音声を聴き取りやすく変換する必要はあるが，発話内容はシステムの前後で大きく異なる必要はない。そこで，2018 年に S. Miyamoto et al.が提案した Pix2Pix を用いたスペクトログラム（音声信号を時間-周波数分析した画像に近いデータ）を話者間で変換する手法の応用を検討した。しかし，先行研究では，このモデルを学習する学習音源数が 100 文と 500 文という極端な値のみであった。我々の提案では，ユーザが事前に音声登録をする場合に利便性が低下するため，最適な学習音源数を比較検討することとした。

我々の提案システムにおける学習フローを図 2 に示す。縦横が異なるが，図 1 の音声合成部モデル部分（図 1 上段）に相当する。元話者はシステムに入力する発話者を指し，目的話者が放送される発話者になる。S. Miyamoto et al.と同様に低品質な音声スペクトログラムをまず生成し，スペクトログラムを高音質化するモデルを後段に配置する。実験に用いる音声の設定数を表1に示す。音源は同一の話者が ATR 音素バランス文（503 文）と日本語音声データベース内の小論文朗読音声（519 文）を朗読している ATR 日本語音声データベースから女声（FKN）と男声（MHT） を選択し，表 1 の文章数で 2 つのモデルの学習と聴取実験に用いるテストデータを分割した。音声の分割のイメージは図 3 に示す通りである。

使用用途	音源	音源数	話者
音響特徴量変換 モデルの学習データ	ATR 日本語音声データベース　のセット B (音素バランス文の発話音声) の A~I セット	450 * 2	FKN,MHT
スペクトログラム高精細化 モデルの学習データ	ATR 日本語音声データベースのセット B (音素バランス文の発話音声) の A~I セット +ATR 日本語音声データベースのセット D (小論文の朗読音声)	969 * 2	FKN,MHT
テストデータ	ATR 日本語音声データベースのセット B (音素バランス文の発話音声) の J セット	53 * 2	FKN,MHT

表 1: 実験条件

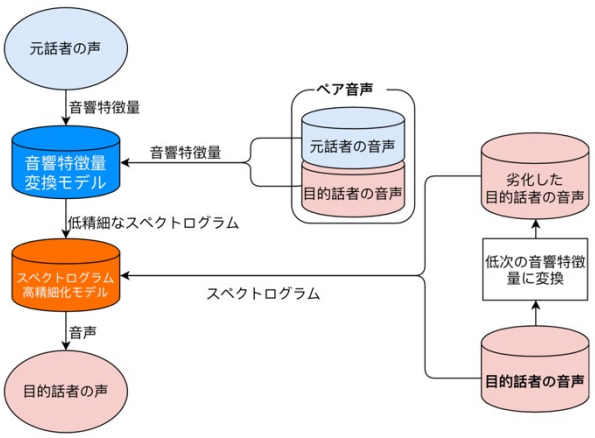

図 2: 我々の実装フロー図

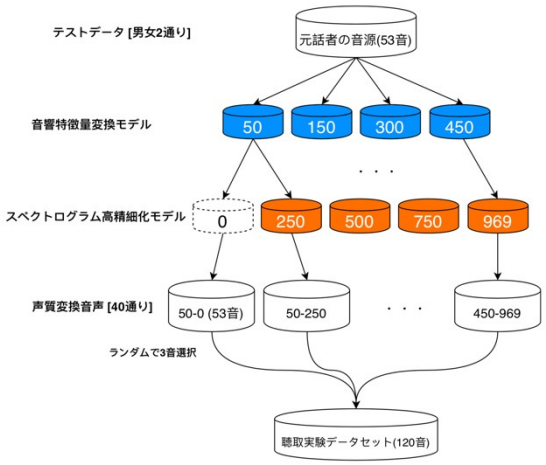

図 3: 我々の実験の音源分割

3. 研究の結論、今後の課題

ここでは声質変換モデルの品質評価実験の結果と課題を述べる。品質評価は5点満点で音声の品質を評価する主観評価方であるMOS (Mean Opinion Score)評価を用いた。被験者は 10 名で，男女話者と学習音源数を考慮した 120 文章について評価してもらった。その結果，低品質モデルの学習に用いる音源数が 150 文章以上高品質モデルの学習に用いる音源数が 750 文章のときMOSの平均値が 3.5 を超える品質の変換音声が得られた。この結果で着目すべき結果は，高音質モデルは最大の学習である 969 文章より少ない 750 文章に品質のピークが見られた点である。低品質モデルも 300 文章が良い結果であることを考慮すると，Pix2Pix 声質変換で用いる 2 つのモデルは主観的な品質面からは学習データに過学習しやすいということが明らかとなった。

本研究で行なった声質変換による拡声システムは，「入力に発話した音声がシステムから拡声されない」という点で，直感と異なる動作をする。この直感的な動作も，電話が発明された(特許取得された)1876 年以降に人類に身についた感覚である。意味情報の伝達という本質に立ち返った時に，元の発話者の音声を伝送すること自体にどれだけの意味があるかを考え直すということを投げかけることのできるシステムとして実装していきたいと考えている。

4. 成果の価値（とくに判りやすく書いて下さい）

4．1．社会的価値

本研究の本質は，拡声システムや電話（音声チャット）によるコミュニケーションにおいて，意味伝達しやすい音声とは何かを考え，実装することである。これにより，聴き返しのない音声コミュニケーションを目指すことが可能であり，特に業務で発生するコミュニケーションを効率化することで，社会全体の生産性向上に寄与することが可能となる。現段階では，人間らしい別の人の声に声質変換するためのデータ数評価にとどまっているが，最終的には誰もが使いやすいシステムとして実装していきたいと考えている。

4．2．学術的価値

本研究は実システムを志向した声質変換を組み込んだ音声システムの実装に関する研究である。3 で述べたように，このシステムの完成形は「入力に発話した音声がシステムから拡声されない」直感と異なるシステムである。このようなシステムを人間が利用する際にどのようなことに着目するかを考えることは認知科学分野に貢献する。また，実システムを志向して学術的な先端技術を実装することは，現実の問題を学術界に提起することが可能ということであり，新たな研究分野につながると考えられる。

4．3．成果論文（本研究で得られた論文等を年代順に書いて下さい。未発表のものは公表予定を書いて下さい）

[1] Hirokazu Akadomari, Yuhi Sato,and Yosuke Kobayashi, “Comparison of the Number of Training Data for Pix2Pix Voice Conversion System,” Proc. The 8th IEEE Global Conference on Consumer Electronics (GCCE 2019), Osaka, Japan, 2019. **accepted**

[2] Hirokazu Akadomari, Kosuke Ishikawa, Yosuke Kobayashi, Kengo Ohta and Junichi Kishigami, “HMM-based Speech Synthesizer for EasilyUnderstandable Speech Broadcasting,” Proc. The 7th IEEE Global Conference on Consumer Electronics (GCCE 2018), pp. 714 – 715, Nara, Japan, 2018. DOI: 10.1109/GCCE.2018.8574710

[3] 赤泊寛和，石川耕輔，太田健吾，小林洋介，岸上順一，”「よく聴こえる」拡声システムのための特定話者に適合した HMM 音声合成システムの評価，” 日本音響学会 2018 年秋季研究大会，1-R-40(Poster), pp.1183-1184，大分県大分市, (2018 年 9 月 12-14 日)

[4] 赤泊寛和，石川耕輔，太田健吾，小林洋介，岸上順一，"HMM 音声合成と発話構文解析を利用した「よく聴こえる」拡声システム," 情報処理学会研究報告, Vol.2018-MUS-119 No.51, 4 pages，東京都文京区, (2018 年 6 月 16-17 日)

文献[1]が声質返還に関する論文，文献[2-4]がシステムのプロトタイピングに関する論文
文献[1]の学会投稿版を本報告書に添付する。