Vol.2018-MUS-119 No.51
Vol.2018-SLP-122 No.51
2018/6/17

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

赤泊 寛和[1,a)]　石川 耕輔[1]　小林 洋介[1,b)]　太田 健吾[2]　岸上 順一[1]

概要：我々は，避難誘導に用いる構文解析を利用した聴き取り改善システムを提案している．本稿では，提案システムに音声特徴分析を加え，アナウンスに発話者の話者性を反映するように，HMM 音声合成システムを実装し，話者性の保持に必要な発話数を調べる主観評価を行った．話者性一致評価では，10 人の被験者で，最高 40% の一致率であり，高い値とはならなかった．このため追加で ITU-R BS.1534-1 勧告の MUSHRA を利用した主観評価を 12 名の被験者に対して行ったところ，学習音声は 50 文でも提案システムの品質は保持できることがわかった．

1. はじめに

屋外拡声機器は，災害時など非常放送に利用される重要な機器であるが，放送音声が聴き取りにくい．東日本大震災では，防災行政無線による避難呼びかけに対し「何か言っていたが，聞き取れなかった」や「何か言っていたが，覚えていない」，「呼びかけはしていないと思う (聞いていない)」との回答が全体の 44% に及んだ [1]．これを受け，日本音響学会にて拡声器の品質基準案 [2] が制定され，既存システムの性能評価・改良が進んでいる (例えば [3] など)．Arai *et al.* は，発話の明瞭性改善のために母音による残響のパワーを低減させ，高明瞭化する手法を報告している [4]．しかし，入力された発話音声自体がボソボソと非常に聴き取りにくい発話であった場合，信号処理による改善には限界がある．

この問題を解決するため，我々は入力された音声を構文解析することで伝えるべき意味情報を担保するシステムを提案した [5]．このシステムは，入力音声を自動音声認識によってテキストに変換し，自然言語処理による構文解析をすることで，よりシンプルで意味を理解しやすい文章に変換し，再度音声合成して放送する．本論文では，このシステムの改善に Hidden Markov Model(HMM) 音声合成システムを実装し，特定話者に適合を行ったので報告する．

[1] 室蘭工業大学
Muroran Institute of Technology, Mizumoto 27-1, Muroran 050-8585, Japan
[2] 阿南工業高等専門学校
National Institute of Technology, Anan College,265 Aoki, Minobayashi, Anan, Tokushima 774-0017, Japan
[a)] 18043002@mmm.muroran-it.ac.jp
[b)] ykobayashi@csse.muroran-it.ac.jp

2. 提案システム

2.1 システムフロー

本研究ではケーススタディとして，「地点 A から地点 B へ避難」と放送することを事例として構文解析を行う．提案システムのフローを**図 1** に示す．図 1 の上部が本研究で改善した HMM 音声合成システムのフローであり，下部は従来システム [5] 全体のフローである．このシステムは大きく音声認識部・構文解析部・音声合成部の 3 つで構成している．

音声認識部では，自動音声認識を用いて入力音声をテキストに変換する．音声認識には大語彙音声認識システムの Julius[6] を module モードで用いた．音響モデルはトライフォンによる DNN-HMM であり，言語モデルは単語 N-gram である．構文解析部では，自然言語処理を用いて，発話内容のテキストより避難元と避難先となる地名を確定し，地名を入れた定型文を生成する．

構文解析部では，形態素解析に MeCab[7]，係り受け解析には CaboCha[8]，長単位解析には Comainu[9] を用いた．構文解析の手順を次に示す．はじめに，形態素解析を用いて自動音声認識によって正しく認識することが出来たフィラー等，放送に不要な部分を除去する．次に，係り受け解析によって避難元と避難先の地名を抽出する．係り受け解析では，**表 1** の避難単語を含む文節と係り受けの関係にある文節を抽出する．次に抽出した文節内の表 1 の避難元単語・避難先単語から避難元の地名か避難先の地名かを判断する．この際に，判断できない地名単語があった場合は後に示す処理によって選択する．次に，長単位解析によって抽出した単語が複合語になるか確認し，必要があれば連結

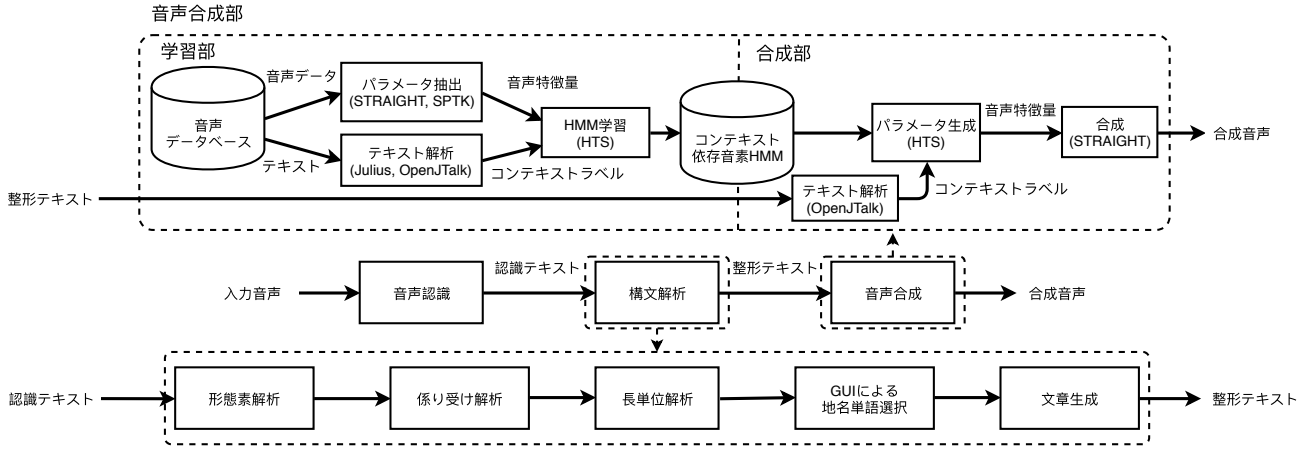

図 1 提案システムのフロー; 図の上部が提案する音声合成システム

表 1 地名単語抽出に用いた単語

意味	単語
避難元単語	より, から, は
避難先単語	に, へ
人単語	人, 方, 皆さん, みなさん
避難単語	避難, ひなん, 非難[*1]
接続詞単語	または, もしくは

する．次に，前述の処理によって抽出した地名単語を確認し，間違いのない放送のため，GUI を用いて人の手で選択する．次に，抽出された情報を元に予め設定した**表 2** の定型文を用いて文章を生成する．最後に生成した定型文から音声合成し，放送のために再生する．地名単語は，日本郵政が公開する郵便番号データを用いて Julius，Mecab，CaboCha，Comainu 用の辞書を作成した．

2.2 システムの改善点

従来システム [5] では OpenJTalk[10] を音声合成に用いた．しかし，OpenJTalk では，配布されている特定の話者モデルを用いて合成するため，マイクロホン入力音声とスピーカ出力音声の音色が一致せず，屋外拡声システムとしては問題がある．この問題は入力話者の音声に適応したモデルの TTS を利用することで解決できる．本研究では，音声合成部に HMM-based Speech Synthesis System (HTS) [11] を用いた音声合成システムを実装し，入力話者の性質を保持して放送するシステムとしたので報告する．

2.3 HMM 音声合成

学習パートでは，音声分析と HMM 学習を行う．テキスト解析部では，テキストからコンテキストラベルを作成する．パラメータ抽出部では，音声信号を分析し，基本周波数 (F_O) とメルケプストラム係数，非周期性指標 (Ap.) を得る．HMM 学習部では，コンテキストラベルと F_O，メルケプストラム係数，Ap. からコンテキスト依存音素 (quin-phone)HMM 学習を行う．

合成パートでは，ラベルの作成と学習済みモデルから音声合成を行う．テキスト解析部では，テキストからコンテキストラベルを作成する．パラメータ生成部では，学習済みのコンテキスト依存音素 HMM とコンテキストラベルから F_O とスペクトラム，Ap. を生成する．音声合成部では，各パラメータから合成音声を生成する．本研究では，入力信号とデータベースから取り出す音声信号は同一話者とすることで，入力音声の話者と合成音声の話者を一致させた．

2.4 HMM 音声合成システムの実装

学習パートについて解説する．パラメータ抽出部では，STRAIGHT [12] と Speech Processing Tool Kit (SPTK) [13] を用いて音声分析を行い，学習に用いるパラメータ分析を行う．GNU Octave を用いて STRAIGHT を実行し，F_O と Ap. 分析を行う．また，SPTK によってメルケプストラム分析を行う．テキスト解析部では，コンテキストラベルの作成を行う．Julius を用いて音素セグメンテーションを行う．OpenJTalk を用いてテキストから生成したコンテキストと音素セグメンテーション結果を結合し，コンテキストラベルとして用いた．HMM 学習部ではコンテキスト依存音素 HMM の学習を行う．HTS を用いて各パラメータと対応するコンテキストラベルから HMM 学習を行い，コンテキスト依存音素 HMM を生成する．

合成パートでは，ここまでの学習パートのおよそ逆の流れで処理を行う．

3. 実装システムの性能評価

3.1 主観評価の概要

新たに実装した音声合成システムによって生成された合成音声の話者性を評価する話者性一致評価，品質を評価する ITU-T BS.1534-1 勧告 MUSHRA(MUltiple Stimulus

[*1] 既報 [5] の解析の際に「避難」を「非難」と誤認識することが多かったため「非難」が含まれている．

表 2　定型文の例

避難元単語の個数	避難元単語の個数	例文
1	1	A の皆さんは、B へ、避難してください。
2	1	A、または、A' の皆さんは、B へ、避難してください。
1	2	A の皆さんは、B、または、B' へ、避難してください。
2	2	A、または、A' の皆さんは、B、または、B' へ、避難してください。

表 3　音声分析条件

サンプリング周波数	16 kHz
量子化ビット数	16 bits
FFT 長	2048
フレームシフト	5 msec.
F_O 範囲	50-670 Hz

test with Hidden Reference and Anchor)[14] を用いた評価をそれぞれ行った．

3.2　録音条件

我々は，HMM 学習に用いる音声として，先行研究 [5] で録音に参加した 20 代男性 1 名に ATR 音素バランス 503 文 [15] を発話してもらい，防音ブース内で録音した．録音した音声のフォーマットは，サンプリング周波数 48 kHz，量子化ビット数 16 bit，リニア PCM，モノラルチャンネルである．録音音声を聴取実験に用いる際には，サンプリング周波数 16 kHz にリサンプリングした．

3.3　音声合成の設定

学習音声データのパラメータとして，メルケプストラム係数を 40 次元と対数 F_O を 1 次元，Ap. を 5 次元，及びそれらの Δ，Δ^2 の計 138 次元の特徴ベクトルを**表 3** の条件で STRAIGHT と SPTK によって分析した．学習に用いる文章数を $50, 100, 150, \ldots, 450, 503$ 文として 10 種類のモデルを作成し，主観評価で比較する．

3.4　MUSHRA

MUSHRA は符号化オーディオの品質評価法である．この評価法は，明示した基準音に対し，隠れ基準音と隠れアンカ音を含めた複数の試験音の劣化量を「非常に良い (Excellent)」に相当する 100 点 (満点) から「非常に悪い (Bad)」に相当する 0 点の間の連続量として評価する．評価対象音に隠れ基準音が含まれていることが明示され，隠れ基準音に相当すると思われる試験音に対し，必ず 100 点を含めるよう被験者に指示する．隠れアンカは，評価の下限値を規定するために含まれており，通常 3.5 kHz をカットオフとするローパスフィルタに適応した試験音を用いる．

MUSHRA は合成音声を評価するための評価法ではない．しかし，合成音声は人間の発話に比べ劣化していると考えることができるため，本研究では MUSHRA を用いた．

3.5　主観評価条件

本研究では，既報 [5] で録音した音声を基準音として話者性一致評価と MUSHRA 評価を行なった．音声はすべて被験者の前 93 cm に設置したラウドスピーカ (YAMAHA MONITOR SPEAKER MS101 Ⅲ) で提示し，音量はリファレンス音源が 60 dB になるよう調整した．提示した評価音声には，3.3 節の条件で生成した合成音声を含めた．話者性一致評価では，既報 [5] で録音した 10 種類の避難放送音声から 3 種類を用いた．被験者は 20 代学生 10 人である．被験者に基準となる録音音声と合成音声を 1 つづつ提示し，話者が一致すると感じるかを評価した．

MUSHRA テストでは，既報 [5] で録音した 10 種類全てを用いた．被験者は 20 代の学生 12 人である．実験 GUI として Web Audio Evaluation Tool [16] を用いて評価を行った．

カットオフ周波数を 3.5 kHz とするローパスフィルタをかけた音声では，HMM 合成音声と同程度でアンカとして不十分なため，追加のアンカとして録音音声にそれぞれ 1.0 kHz と 2.0 kHz, をカットオフ周波数とするローパスフィルタをかけた音声も含めた．

3.6　話者性一致評価の実験結果と考察

表 4 に話者性一致評価の音声の種類ごとのスコアを示す．スコアは，話者が一致する場合を 1 点，一致しない場合を 0 点とし，学習条件ごとの平均値とした．L50 から L503 は，学習音声数ごとの合成音声を表す．

話者性一致評価の結果から，学習に用いる音声の数に寄らず，話者性の一致率が低いことがわかる．また，学習に用いる音声数ごとにスコアに上下がある．学習に用いた録音音声は，プロではない 20 代男性による読み上げ音声であるため，データセット内で発話条件が一致していない可能性がある．1 日あたり 100 文程度のペースで行ったため，録音日によって基本周波数の上下あるなど，聴取感に差がある．そのため，合成音声と録音音声を比較した際に，学習に用いた文章セットによっては話者性が一致していないと感じる傾向にあると考えられる．

3.7　MUSHRA テストの実験結果と考察

図 2 に MUSHRA テストの評価スコアを箱ひげ図で示す．L50 から L503 は，話者一致評価と同様である．anc1.0 と anc2.0, anc3.5 は，それぞれアンカである．ref. は隠れ

表 4　話者性一致評価のスコア

合成音声	L50	L100	L150	L200	L250	L300	L350	L400	L450	L503
スコア	0.27	0.27	0.33	0.33	0.2	0.23	0.4	0.27	0.27	0.33

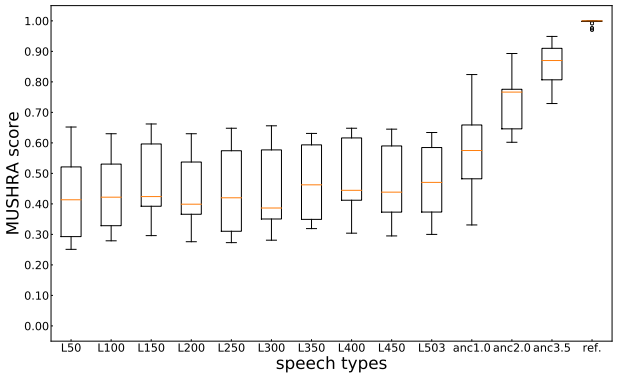

図 2　MUSHRA の結果

基準音である．

MUSHRA テストの結果より，合成音声は，学習音声数に寄らず，ばらつきに差がないことから，合成音声のスコアは，学習音声数に依存しないことがわかった．このことから実装した音声合成システムでは，合成音声の品質は 50 文の学習でも十分な品質である．アンカに比べスコアが低い理由として，話者性一致評価と同様に，基本周波数の上下があることがあることが挙げられる．そのため，学習音声数を増やしてもスコアの向上に繋がらなかった可能性があり，改善する必要がある．

4. まとめ

我々は提案する「わかりやすい」屋外拡声システムの改良に HMM 音声合成を用いた．実装した音声合成システムを話者性一致評価したところ，話者性において課題が残った．また，MUSHRA によって被験者評価したところ，合成音声の品質は，学習音声数に依存しないことがわかった．今後，特定話者以外の音声入力で話者性を合成音声に反映させるため，話者適応や声質変換手法の検討を行う．

実装の課題として，テキスト解析に OpenJTalk のみを用いていることがある．学習音声のアクセントなどのラベルが正確ではない可能性があり，より良い実装方法について検討する必要がある．

謝辞　本研究を遂行するにあたり，HMM 音声合成システムについてのアドバイスをいただいた山形大学の小坂哲夫教授，相澤佳孝さんに深く感謝する．また，本研究の一部は JSPS 科研費 (16K21584)，(公財) 人工知能研究振興財団，(公財) 電気通信普及財団，(公財) 国際科学技術財団，(公財) 立石科学技術振興財団，東北大学電気通信研究所共同研究プロジェクト (H29/A18) の助成を受けた．関係者と被験者各位に感謝する．

参考文献

[1] 内閣府, “東北地方太平洋沖地震を教訓とした地震・津波対策に関する専門調査会, 第 7 回会合”, p. 54, 2011.
[2] 日本音響学会, “災害等非常時屋外拡声システム性能確保のための ASJ 技術規準”, 2017.
[3] Taira Onoguchi and Yoshifumi Chisaki, “Emission timing controller by single board computer for public address system,” Proc. 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE), pp. 315–316, 2016.
[4] T. Arai and K. Kinoshita and N. Hodoshima and A. Kusumoto and T. Kitamura, “Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments,” Acoustical Science and Technology, vol. 23, no. 4, pp. 229–232, 2002.
[5] 石川耕輔, 太田健吾, 小林洋介, 岸上順一, “発話構文解析を利用した「よく聴こえる」拡声システムの基礎検討”, 研究報告音楽情報科学, vol. 2017-MUS-15, no. 37, jun 2017.
[6] Julius project team, “Julius”, [Online]. Available: http://julius.osdn.jp
[7] 工藤拓, “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト. [Online], Available: http://taku910.github.io/mecab/
[8] T. Kudo and Y. Matsumoto, “Japanese Dependency Analysis using Cascaded Chunking,” CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002, 2002, pp. 63–69.
[9] K. Uchimoto and Y. Den, “Word-level Dependency-structure Annotation to Corpus of Spontaneous Japanese and Its Application,” LREC 2008: Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008, pp. 3118–3122.
[10] HTS working group. “Open JTalk”. [Online]. Available: http://open-jtalk.sp.nitech.ac.jp
[11] HTS working group and others. “HMM/DNN-based Speech Synthesis System(HTS) - Home”. [Online]. Available: http://hts.sp.nitech.ac.jp
[12] Kawahara, Hideki and Masuda-Katsuse, Ikuyo and De Cheveigne, Alain, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1,” Speech communication, vol. 27, no. 3-4, pp. 187–207, 1999.
[13] SPTK working group. “Speech Signal Processing Toolkit (SPTK) Version 3.11 December 25, 2017”. [Online]. Available: http://sp-tk.sourceforge.net
[14] ITU, “Method for the subjective assessment of intermediate quality level of coding system,” 2003.
[15] 匂坂芳典, 浦谷則好, “ATR 音声・言語データベース”, 日本音響学会誌, vol. 48, no. 12, pp. 878–882, dec 1992.
[16] Jillings, Nicholas and Moffat, David and De Man, Brecht and Reiss, Joshua D., “Web Audio Evaluation Tool: A browser-based listening test environment,” in 12th Sound and Music Computing Conference, July 2015.