Vol.2017-MUS-115 No.37
2017/6/17

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

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

概要：避難誘導等で用いられる屋外拡声器の放送音声は，聴き取りにくいことが指摘されている．本稿では，構文解析を利用したテキスト処理により避難誘導放送を簡潔でわかりやすい文章にして放送するシステムを検討した．提案システムでは，発話音声を音声認識によりテキストを抽出し，構文解析により必要な情報を抽出した後に放送用音声に合成して放送する．特に，構文解析では，係り受け解析を用いることで避難誘導放送のための避難元地名と避難先地名を抽出し，定型文に当てはめ，発話された地名情報を保持した放送文章を作成する．本システムの評価実験で元の音声と合成音声の意味比較の評価実験より，一致率が 79.0 %の結果となった．

1. はじめに

同報系防災行政無線の屋外拡声器（以下，拡声器）は，東日本大震災など多くの災害において避難誘導に利用される重要な機器である．一方，拡声器の放送音声は品質が悪いことも指摘されている [1]．これを受け，日本音響学会にて拡声器の品質基準案が制定 [2] 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）．我々の研究グループでもこれまでに音声認識を用いた品質評価について検討してきた [4]．

拡声音声の聴こえについて Arai らは，発話の明瞭性改善のために拡声する音声の定常部を抑圧する信号処理を行うことで，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 [5]．この手法は，既存のシステムが出力する音声自体を信号処理によって加工する．このようなシステムは，元の音声の話者性を損なう可能性があるが，非常用放送を想定した場合，発話者の話者性を担保する必要は少なく，伝えるべき意味情報が保持されていれば，その目的は達成されると考えられる．しかし，元の発話音声自体がボソボソと聴き取りにくい発話であった場合の改善には限界がある．

そこで，我々は，伝えるべき意味情報の保持を元の発話音声への信号処理だけでなく，一旦テキストに変換したうえで音声に含まれる意味情報を保持した簡潔な文章だけを放送するシステムを提案する．提案システムは，入力音声の音声認識によって得られたテキストを構文解析することで，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．これにより，誰が聴いても意味が伝わるシステムとなることが期待される．本稿では，提案システムのプロトタイプを実装したので報告する．

[1] 室蘭工業大学
Muroran Institute of Technology
[2] 阿南工業高等専門学校
National Institute of Technology, Anan College
[a)] 17043002@mmm.muroran-it.ac.jp
[b)] kengo@anan-nct.ac.jp
[c)] ykobayashi@csse.muroran-it.ac.jp
[d)] jay@csse.muroran-it.ac.jp

2. システムの概要

2.1 本システムの概要

提案システムが想定する緊急時の放送音声として，

「水元から中島または登別へ避難してください」

という文章を考える．緊急時に基地局でこの文章の読み上げを想定すると，発話者の焦りなどから文章通りに落ち着いて発話しにくく，発話文に「えっと」等のフィラーや言葉の言い換えが発生しやすい．

「えっと…水元から中島
または登別の方へ避難してください」

この文章では，端的に情報を伝えることができず，聴き取りにくさより情報伝達を阻害する．そこで，

「A の皆さんは，B へ避難してください．」

という必要最小限の情報のみからなる定型文に変換する．これにより，フィラーが無く端的に避難対象地域と避難先地域を提示し，より聴き取りやすい文章になる．

本研究では，例示した避難元と避難先からなる事例の発話のみに限定して，提案システムのプロトタイプを実装し，放送内容の意味を保持した拡声システムが実現可能か検証

する．将来的には一般の屋外拡声システムを想定した様々な放送が可能なシステムに発展させていく．

今回作成したシステムの全体図を**図 1** に示す．本システムは主に音声認識部・構文解析部・音声合成部の 3 つの主要ブロックで構成した．入力音声を音声認識部でテキストに変換する．構文解析部では，発話された地名を確定し，地名を入れた定型文を生成する．最後に生成した定型文で音声合成し，再生する．合成した音源はファイル保存するので，ループ再生も可能である．これにより，同じ文章であれば避難勧告文を何度も読み上げる必要が無くなり，言い間違いの回数が減り，誤情報を流さずに済む．

2.2 音声認識部

音声認識部では，入力された音声をテキストに変換するために，大語彙音声認識システムの Julius[6] を module モードで用いた．本研究では，利用対象である室蘭の地名の認識精度を向上させるために辞書に地名を辞書に追加した．辞書については 3.1 節で詳しく説明する．この他の設定は Ver. 4.4.2 の初期設定パラメータをそのまま用いたため，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．

2.3 構文解析部

図 1 の構文解析部の詳細を示す．まず，室蘭市の地名の辞書を追加した MeCab[7] を用いて形態素解析を行い，次に CaboCha[8] を用いて係り受け解析をする．ここで係り受け情報と地名辞書から 2.1 節の例文における地名単語 A と B を抽出する．取得した地名単語は Comainu[9] による長単位解析を行い，複合語になるか解析し，必要があれば連結する．

抽出された地名単語は係り受け情報を参照して，避難元の From 側の地名と，避難先の To 側の地名，及び解析からは判別できずにどちらにも入れられなかった地名である Unknown の 3 パターンに分類する．地名単語の分類には，避難誘導文を扱うため「避難」という単語に着目し，「避難」を含む文節と係り受け関係にある文節を抽出する．係り受け関係にある文節が 2 つの場合は**表 1** に示す文節内の格助詞の組み合わせで From と To のどちらになるかを判断する．どちらにも当てはまらない場合は Unknown として処理する．「A または B」のように複数の地名単語が同時に入力されると，複数の地名単語が連結した係り受けとなる．この場合は繰り返し係る文節を全て抽出して From または To に追加する．係り受け関係にある文節が 2 つ以外の場合は，係り受け解析か音声認識処理に問題があると判断し，認識した文節に含まれる地名単語を Unknown として表示する．

提案システムは防災用であり，誤情報の放送を行うわけにはいかない．よって，誤った放送を無くすための最終的な確認を人間が目視で確認できるようにする．このために確認表示用 GUI を作成した．GUI 部では，From 地名と To 地名には，解析結果を基にした地名単語を表示し，Unknown は両方に表示し，最終的に人間がどちらに入るか判断する．

表 1 List of case-marking particles

Mean	Particle
From location	より，から
To location	に，へ

2.4 音声合成部

音声合成には，OpenJTalk[10] のうち Node.js で利用可能なモジュール [11] を GUI での操作を行うために利用して合成音声を作成し，一旦ファイル保存してから再生する．合成用のパラメータは，3.3 節に記述するチューニングを行い，品質を調整した．

3. システムのチューニング

3.1 辞書の作製

本システムでは，放送するエリアに含まれる地名を正しく認識する必要がある．例えば，室蘭市にある「蘭北（らんぽく）」のような地域固有の名詞は，標準的に用いられる既存の辞書を用いた場合に認識することが出来ない．そこで，室蘭市が提供しているオープンデータ [12] から室蘭市のバス停名と避難場所名を，日本郵政が公開する郵便番号のデータ [13] から室蘭市の住所地名をそれぞれ取得した．取得した上記の名詞データを Julius，MeCab，OpenJTalk 用の辞書に対応したフォーマットに成形し，それぞれの辞書に追記した．OpenJTalk 用辞書にはアクセント情報も追記した．

3.2 音声認識部のチューニング

3.2.1 概要

本システムでは，入力音声から定型文に挿入する地名を正確に取得する必要がある．フロントエンドに利用する音声認識部の認識性能はシステム全体のユーザビリティを決定するため，想定文章の音声を作成し，認識性能の評価を行った．

3.2.2 性能評価音源

音声認識部の性能評価音源には，作成した想定文章を 10 名の参加者に読み上げてもらい，その音声を録音した．この音声を，Julius で認識する際の認識精度を検証した．認識精度の評価と並行し，3.1 節で追記した辞書の性能を確認するため，Julius の Dictation-kit の既存辞書のみを利用した場合と，室蘭市の地名が追加された辞書を利用した場合の 2 つを比較した．本評価で用いる文章を**表 2** に示す．この文章は，作成した地名辞書により 2.1 節に示した例文

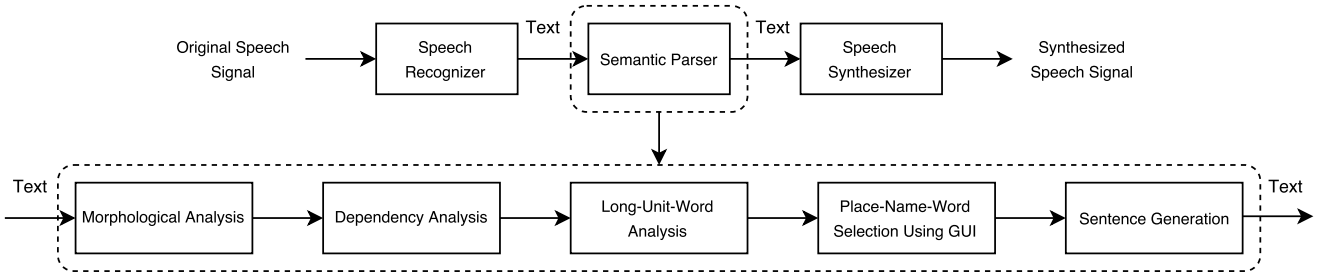

図 1 Proposed System flow

表 2 Example sentence

Number	Sentence(Japanese)
1	蘭北団地から千歳町へ避難してください
2	文化センターから中島公園へ避難してください
3	文化センターから高砂公園へ避難してください
4	製鉄記念室蘭病院からすずかけ地蔵尊へ避難してください
5	室蘭観光協会から工大へ避難してください
6	室蘭築港から工大へ避難してください
7	知利別郵便局から中島公園へ避難してください
8	祝津町から工大研修所へ避難してください
9	室蘭養護学校から緑町へ避難してください
10	労働福祉センターから緑丘へ避難してください

の地名単語 A と B をそれぞれ 1 単語ずつ抽出し 10 文生成した．

評価文の発話者には 20 代男性 10 名に参加してもらった．被験者は録音前に一度だけ，例文を読む練習を行ってから室蘭工業大学に設置した防音ブース内で音声録音した．録音のフォーマットはサンプリングレート 48 kHz，量子化ビット数 16 bit のリニア PCM である．発話者より明らかな読み間違いがあったと申し出があった発話については，同一話者で再度録音を行った．

3.2.3 評価指標

評価結果を式 (1)〜(3) の適合率, 認識率及び F 値を Julius 辞書のみと地名辞書を追加した辞書のそれぞれで求めた．

$$\text{適合率} = \frac{\text{正しく抽出できた地名数}}{\text{抽出できた地名数}} \quad (1)$$

$$\text{再現率} = \frac{\text{抽出できた地名数}}{\text{抽出できる地名の総数}} \quad (2)$$

$$F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{(\text{適合率} + \text{再現率})} \quad (3)$$

3.2.4 認識結果

認識結果の各指標を**表 3** に示す．結果より，辞書を追加した場合に F 値が 0.251 向上し，0.910 になった．このことから，提案システムのように特定の地域のみで利用することを想定したシステムでは，地名の辞書を追加することにより高精度な認識が可能になることがわかる．一般的な大語彙音声認識では大規模コーパスにより名詞を取得する必要があるが，提案システムのように特定の地域のみで利用されることを想定すると，本システムのような小規模な追加で十分であることがわかった．

表 3 Speech recognition result

Dictionaly set	Precision	Recall	F value
Default dictionaly	0.683	0.636	0.659
Default dictionaly + Place-Name-Word dictionaly	0.916	0.904	0.910

表 4 MOS rating category

Category	Score
非常に良い	5
良い	4
普通	3
悪い	2
非常に悪い	1

3.3 音声合成の設定

3.3.1 概要

本システムの最終出力では，合成音声を再生するため，聴き取りやすい音声を合成する必要がある．OpenJTalk には合成品質を調整するパラメータが設定されている．本研究では，複数の合成パラメータの組み合わせを主観評価によって比較し，最適な値を決定した．主観評価には MOS(Mean Opinion Score)[14] による 5 段階評価を行った．

3.3.2 評価指標

音声評価は，MOS による主観品質評価で行った．MOS は，実体範疇尺度法 (ACR: Absolute Category Rating) を用いた音声の主観評価尺度である．使用した評価指標の一覧を**表 4** に示す．評価には，専用 GUI を作成した．比較した 125 条件のうち，全被験者の平均値が最も良いパラメータの組み合わせを最適値とした．

3.3.3 評価手順

評価の被験者には 20 代男女 12 名に参加してもらった．評価は室蘭工業大学に設置した防音ブース内でオーディオインターフェース (Roland UA-25EX) に接続したヘッドホン (Senheizer HDA-300) を装着して行った．合成音声の発話文章は，ATR 音素バランス 503 文 [15] より文章を選択して利用した．作成した音源のフォーマットはサンプリングレート 48 kHz，量子化ビット数 16 bit のリニア PCM である．被験者には評価音源の聴取前に操作説明のために本番では用いない 5 文章とパラメータで作成した合成音声を用いて GUI の操作練習を行いながら，評価音源の再生レベ

表 5　OpenJTalk setting parameters

Parameter	Value
All-pass constant	0.3, 0.4, **0.5**, 0.6, 0.7
Additional half-tone	-10.0, -5.0, **0**, 5.0, 10.0
Weight of GV for log F0	0.1, 0.5, 1.0, 2.0, **3.0**

ルをちょうどいい音量になるように決定してもらった．評価音源には，**表 5** に示したパラメータの組み合わせ 125 パターンより 125 文を生成して利用した．被験者は操作練習の際に最適に設定した音量に固定して MOS 評価を行った．

3.3.4　評価結果

全 125 の評価音源のうち，最も評価が高いパラメータの組み合わせを表 5 の太字部分で示す．このときの MOS 評価値は 4.58 であり，次点の値より 0.08 高かった．なお全体の平均値は 2.99 で，標準偏差は 0.74 と，設定した値は多くの組み合わせよりも高品質であった．この設定値を用いて提案システムでの放送音声を合成する．

4.　システム全体の性能評価

4.1　概要

各部をチューニングし，完成した提案システム全体の品質評価を行う．本稿では，音声認識部の性能評価に用いた音声より生成したテキストを処理し，最適値にチューニングした合成音声の聴取感を評価する．本稿では，図 1 の提案システムにおける構文解析部が無い場合は，F 値が 0.910 の音声認識であるため，発生する誤認識単語により入力音声と合成音声との意味が変わってしまうが，構文解析部があれば，元の文章の意味をそのまま保持するため，意味が変わらないと考えられる．

4.2　評価方法

本評価では，3.2 節の音声認識性能評価の際に録音した音声と提案システムによる構文解析を用いた合成音声の意味が一致するか比較する．比較のため，構文解析部を用いずに音声認識の結果のテキストから合成した音声も比較する．これにより構文解析の効果を検証する．評価音源は録音した音声を先行音に，テキストによる合成音声を後続音として結合した評価音源を被験者にランダムに提示し，意味が一致するかしないかの 2 択で強制回答する GUI を作成して評価した．この際に被験者には「これから流れる音声はすべて避難誘導放送です．2 つの音声が流れますので，意味が一致するか異なるか判断して回答してください.」という教示のみを行った．

評価音を作成するための基となる音声は，音声認識の性能評価用に収集した音声であり，被験者 10 名それぞれの 10 発話である．しかし，発話者から申し出のなかった読み間違い音源が 1 つ存在したため，これを除いた 99 発話を用いた．これらの音声を音声認識し，提案システムでテキスト処理を行った．なお，今回利用する表 2 の音声に含まれる地名単語は全て 2 単語である．地名の係り受け解析による選択に誤りが 8 発話あった．加えて，2.3 節で述べた「避難」自体が認識できずに提案システムが入力音声をリジェクトした音声が 22 発話あった．よって，利用した音声の総数は 77 発話であり，構文解析部の有無で 2 倍の 154 音が評価総数となる．

主観評価に参加した被験者は 10 名であり，3.3.3 節と同一の評価環境によるヘッドホン聴取で評価を行った．なお，再生レベルの調整には，3.3.3 節の練習フェーズで用いた音源をそのまま利用した．

表 6　Performance evaluation result

Settings	Matching rate (%)
Without parser	78.8
With parser	79.0

4.3　評価結果

結果を**表 6** に示す．結果より，構文解析部を利用することで 0.2 %向上したが，2 つの結果に対して片側 t 検定を行った際に $t(18) = 0.761, p = 0.456$ であり，有意な差は見られなかった．この要因には，全ての評価音源において人間が発話した音声が必ず先行音であったこと，2 つの文章を記憶して比較するため短期記憶だけでは判断が困難であったことなどが考えられる．よって今後は，意味の一致という抽象的な問いではなく，From と To に相当する地名を選択回答する評価実験などを検討する．

5.　まとめ

本稿では，避難誘導放送をテキスト処理によって「よく聴こえる」拡声システムを提案し，実験によって提案システムの評価を行った．音声認識評価より，地名単語を辞書に追加すると F 値が 0.251 向上し，有用性を確認でき，音声合成の主観評価実験より，最適なパラメータを設定できた．しかし，システム全体の性能評価では構文解析部の有無による差が $t(18) = 0.761, p = 0.456$ となった．

今後の課題として，音声認識部では同音異義語の認識誤りが存在したため対応する必要がある．構文解析部では，本稿で行った評価実験では有意な差が出ないため，被験者への教示文の変更や提示順番等の主観評価実験系の再構築を検討する．

謝辞　本研究を遂行するにあたり，システムの実装に協力して頂いた室蘭工業大学の平森貴嗣君，藤田優貴君に感謝する．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報告 (2011).
[2] 日本音響学会災害等非常時屋外拡声システムのあり方に関する技術調査研究委員会：災害等非常時屋外拡声システム性能確保のための基準案および解説.
[3] Onoguchi, T. and Chisaki, Y.: Emission timing controller by single board computer for public address system, *Consumer Electronics (GCCE), 2015 IEEE 4th Global Conference on*, pp. 315–316.
[4] 小林洋介，太田健吾，近藤和弘：機械学習と音声認識による拡声音声品質予測，情報処理学会研究報告，Vol. 2016-MUS-111, No. 43, pp. 1–4 (2016).
[5] Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A. and Kitamura, T.: 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).
[6] development team, J.: Julius, Julius development team (online), available from ⟨http://julius.osdn.jp/⟩ (accessed 2017-5-1).
[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 from ⟨http://taku910.github.io/mecab/⟩ (accessed 2017-5-1).
[8] 工藤 拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol. 43, No. 6, pp. 1834–1842 (2002).
[9] 小澤俊介，内元清貴， 伝康晴：BCCWJ に基づく中・長単位解析ツール Comainu，言語処理学会第 20 回年次大会論文集，pp. 582–585 (2014).
[10] working group, H.: OpenJTalk, nitech (online), available from ⟨http://open-jtalk.sp.nitech.ac.jp/⟩ (accessed 2017-5-1).
[11] 凹 (id:hecomi)：Node.js TTS module using OpenJTalk, hecomi (online), available from ⟨https://github.com/hecomi/node-openjtalk/⟩ (accessed 2017-5-1).
[12] 室蘭市：むろらんオープンデータライブラリ，室蘭市（オンライン），入手先 ⟨http://www.city.muroran.lg.jp/main/org2260/odlib.php⟩（参照 2017-5-1）.
[13] 日本郵政グループ：郵便番号データダウンロード，日本郵政グループ（オンライン），入手先 ⟨http://www.post.japanpost.jp/zipcode/download.html⟩（参照 2017-5-1）.
[14] NTT: 音声品質評価法，NTT（オンライン），入手先 ⟨http://www.ntt.co.jp/qos/technology/sound/03_1.html⟩（参照 2017-5-1）.
[15] 国立情報学研究所：音声資源コンソーシアム，国立情報学研究所（オンライン），入手先 ⟨http://research.nii.ac.jp/src/ATR503.html⟩（参照 2017-5-1）.