声優統計コーパス：
二次配布可能な音素バランス文とその読み上げ音声の構築

@y_benjo and @MagnesiumRibbon

Abstract

声優統計コーパスについて，特に音素バランス文の選択に着目して説明する．

Keywords: 声優統計コーパス，音素，バランス文，ATR503 文，Wikipedia，Neologd，0-1 整数計画問題

1. はじめに

声優オタクであれば，一度は「声優の演技を自分で収録してみたい」と考えたことがあるはずだ．声優統計の代表である筆者も例外ではない．特に，

- 声優の声は一般人のそれとはどのように異なっているのか
- アルゴリズムは話者を識別，および模倣できるのか
- 好きな声優の声を人工的に作り出せるのか

といった疑問に取り組むためには，声優の発声を録音したもの (以降，本稿では音声の集合を音声コーパスと呼ぶ) の構築が必要不可欠である．

我々声優統計としても，大学や企業に属さず，音声コーパスを持たない野良研究者の助けとなるような，自由に利用可能な音声コーパスを構築し，配布することにより，コミュニティ全体による声優統計活動を活性化したいという意思がある．

この度，我々日本声優統計学会は，サークル活動の集大成として声優統計コーパスを構築し，ウェブサイト [1] にて無償配布を行ったが，その構築は一筋縄ではいかなかった． 本稿では，

- 音声にまつわる研究活動に利用が可能である
- 無料で，自由に配布，および利用が可能である

これらの条件を満たす音声コーパスはいかにして構築可能であるかについて検討する．本稿では特に，音声コーパス構築に必要な音素バランス文の選択について 0-1 整数計画問題による定式化を行う．また，構築したバランス文を台本とした音声収録において得たさまざまな気づきも記す．

本稿が音声コーパスの自作を試みる読者の助けとなれば幸いである．また，ここに記す音声コーパスの作成手続きについては不備なども存在するだろう．識者の忌憚のない意見もいただければ幸いである．

本稿の構成は大きく分けて三つである．まず，今回公開した声優統計コーパスについてその概要を説明し，その後，既知の資源の概要と課題について述べる．その後，声優統計コーパスを構築する手続きについて順を追って説明し，構築や音声収録における気付き，反省について述べる．

本稿は声優統計第九号で発表したもの [3] に加筆修正を行ったものである．もし手元にそちらがある場合は，差分にも着目して欲しい．

2. 声優統計コーパス

声優統計コーパス（以降，本コーパス）はその名の通り，日本声優統計学会が構築した音声コーパスである．

声優統計コーパスは

音素バランス文 日本語版 Wikipedia の本文データをもとに構築した 100 個の音素バランス文

音声ファイル 三名のプロの女性声優が「ニュートラル」,「怒り」,「喜び」の三種類の感情表現で読み上げた 900 個の音声ファイル

から構成されている．

本コーパスの特徴は次の四つである．

- 音素バランス文が CC-BY-SA ライセンスで利用，再配布が可能であること

[1] http://voice-statistics.github.com

- 音素バランス文には日本語版 Wikipedia 本文データに登場する diphone の上位 500 種が全て 3 回以上登場すること
- 音声ファイルは無料で研究・分析目的に利用できること
- プロの声優による感情表現が含まれた音声ファイルであること

本コーパスが想定する利用者層は特に

- 音声コーパスを持たない，趣味で研究や分析を行なう者
- 既存の音声コーパスを持つ専門の研究者

の二種類である．

前者には，高価な音声コーパスを購入する前の分析の足がかりとして，後者には，既存の音声コーパスと組み合わせた感情合成，感情認識などに活用していただければ幸いである．

3. 既知の資源

続いて，単一話者による発話が収録された音声コーパスの中でも最も代表的なものである ATR 音声データベースについて述べる．

3.1. ATR 音声データベース

日本語での音声コーパスにおいて最も著名なものは ATR によるデジタル音声データベースであろう．特に B セットに含まれる 503 文は通称「 ATR 503 文」とも呼ばれ， Twitter bot なども存在している．読者の中には正式名称は知らずとも

> あらゆる現実をすべて自分のほうへねじ曲げたのだ。

という，なんとも奇妙な，どこか心をざわつかせる文章に見覚えがある方もいるのではないだろうか．

ATR 503 文は

> 新聞、雑誌、小説、手紙、教科書等の文献から無作為に抽出した約１万の文をもとに、音素環境をバランスさせて作成した 503 文（音素バランス文）

であり [2]，B セットはこれを複数の話者が読み上げたものが収録されている．音声にまつわるどの研究室にも備えられている，音声認識や音声合成を行う上で最も基本となる音声コーパスである．価格は話者一人あたりにつき 35 万円であり，野良研究者が買うには少し勇気が必要な価格である．

503 文の大きな特徴は，日本語において頻出する 2 音素 (**ダイフォン**) と 3 音素 (**トライフォン**) をバランスよく含むよう設計されている点である．例えば「ドーム」という単語の場合，読みは「d/o:/m/u」となり，この時のダイフォンは d/o:，o:/m，および m/u であり，トライフォンは d/o:/m, o:/m/u である．これらダイフォンとトライフォンがバランス良く含まれているため，日本語の音声研究において ATR 503 文，およびそれを読み上げた B セットがデファクトの音声コーパスとなっている．

3.2. ATR 503 文の課題

では，ATR 503 文を声優に読み上げてもらい，その音声を無償配布すればいいかというと，それは不可能である．

なぜなら，ATR 503 文を読み上げた音声コーパスだけでなく，503 文そのものも ATR によって販売されているためである．これはすなわち，文自体の知的財産権も ATR が有しており，読み上げた音声コーパスの配布は文の配布とほぼ同義であるため，ATR の権利を侵害する行為となるためである [3]．

4. 声優統計コーパスの構築

前章では，音声コーパスとしての ATR デジタル音声データベース B セットと，そこに含まれる ATR 503 文を紹介し，ATR 503 文をそのまま利用した音声コーパスを配布することは権利上難しい事について述べた．そこで本稿では，

- ATR 503 文のように音素 [4] のバランスを考慮し
- 二次配布が可能である

バランス文集合を構築し，その文章を声優に読み上げてもらうことで自由に配布が可能なコーパスである声優統計コーパスを構築する．

4.1. 概要

コーパス構築の手続きは以下のように分解できる．

- まずバランス文の候補となる，二次配布が可能な大量の文章集合 (文候補) を入手する
- その後，文候補の各文にダイフォンを付与する

[2] 上記の文も含め http://www.atr-p.com/products/sdb.html から引用

[3] 検索する限り，前述の Twitter bot や個人のブログにおいても全文が公開されているが，これは著作権侵害である可能性が非常に高い．

[4] 今回はダイフォンまでを考慮する．

- バランス文選択を 0-1 整数計画問題として定式化する
- 定式化した 0-1 整数計画問題を解く
- 解を人手で確認し修正する
- 得られたバランス文を実際に声優に読み上げてもらう

各処理について順を追って説明する．

4.2. 文候補の入手

まず，バランス文の候補となる文 (文候補) を用意する必要がある．文候補は以下の性質を備えていなければならない．

- 大量に用意されていること
- 二次配布が可能であること
- 日本語としてある程度整っていること

一つ目の条件は二つの観点から必要である．まず第一に，我々は「発話される日本語において登場するダイフォンの真の分布」というものを持っていない．これはつまり,「今回構築する音素バランス文においてどのダイフォンまでをカバーする必要があるか」がわからない．今回我々は大量の日本語文章を用意し，その中に含まれるダイフォンを数え上げ，その分布が真の分布に似たものになる，という仮定を採用した．そのためにも，大量の日本語文章が必要となるのである．また，文選択を 0-1 整数計画問題として解くという観点から，文候補が多いほうが (計算コストが増えるとはいえ) より「良い文集合」，少ない文章・文字数で十分にダイフォンをカバーするバランス文集合が得られやすい，というものがある．

v 二つ目の条件については今回構築する音声コーパスの目的からして不可欠である．よって，自然言語処理研究で用いられることの多い毎日新聞コーパス [5] などは文候補として採用できない．

最後の条件は後段の処理に関わるものである.「音素バランス文」と言うからには文章に対して音素，すなわち「読み」を付与する必要がある．さすがに人手で全ての文章に読みを付与するわけにはいかず，既存のソフトウェアを用いるわけだが，そのためにはアルゴリズムが処理しやすいよう，ある程度日本語として整った文章が必要とされる．例えば，ニコニコ動画に投稿されたコメントデータセット [6] は総コメント数約 35 億件と非常に膨大な量ではあるが，崩れた文章が多いために文候補として採用するのは好ましくないと考えられる．

今回我々はこれら三つの条件を満たし，最も入手が容易なデータセットである日本語版 Wikipedia[7] の記事本文全件データ (jawiki-latest-pages-articles.xml.bz2) を用いることにした．

まず一つ目の条件については，現在日本語版 Wikipedia には 100 万件以上の記事が存在しており，十分な量の文が存在していると考えていい．また，Wikipedia データのライセンスは CC-BY-SA であり，二次配布が可能であるため，二つ目の条件についても満たしている．最後に， Wikipedia の文章はボランティアによって編集が繰り返されているため，内容の真偽はさておき，文章が破綻していないものが多く，アルゴリズムによる処理にも耐えうると考えられる．

4.3. 文章へのダイフォンの付与

文候補が定まったところで，文候補に含まれる各文章に対して読みを付与し，その上でダイフォンの集合に変換する必要がある．今回は読み付与に形態素解析ソフトウェアである MeCab[1] を用いる．

簡単のため `~/.mecabrc` に

```
; hatsuon
node-format-hatsuon = %pS%f[8]
unk-format-hatsuon = %M
eos-format-hatsuon  = \n
```

と書くことで，MeCab を用いた読み付与は次のように実現できる [8].

```
% echo ’情報の数量的認識’ | mecab -Ohatsuon
ジョウホウノスウリョウテキニンシキ
```

しかし，MeCab 標準の辞書に含まれない単語については次のように読み付与が失敗してしまう.

```
% echo ’佐倉綾音’ | mecab -Ohatsuon
サクラアヤオン

% echo ’水瀬いのり’ | mecab -Ohatsuon
ミズセイノリ
```

[5] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[6] http://www.nii.ac.jp/dsc/idr/nico/nico.html
[7] https://ja.wikipedia.org/
[8] ここでは読み部分だけを取得しているが，後述の処理において全ての結果を取得する必要があることに注意しよう

そこで，Neologd[2] も合わせて用いることにより，読み付与の精度を改善する．Neologd は様々なデータソースから取得したコーパスにもとづき，新語や未知語を適切に解析するための辞書である．MeCab 用の Neologd をインストールした状態で先ほどの例を実行すると

```
% echo ’佐倉綾音’ | mecab -Ohatsuon
サクラアヤネ

% echo ’水瀬いのり’ | mecab -Ohatsuon
ミナセイノリ
```

といったように改善されていることがわかる．

読みが付与できれば，あとは音素に変換し，ダイフォンの集合へと変換すればよい．今回は読みから音素へのマッピングはオープンソースの音声認識エンジン Julius[9] に含まれる変換規則 [10] に従うことにした．その結果，

```
% echo ’ドーム’ | ruby phonology.rb
phonology : d/o:, m/u

% echo ’佐倉綾音’ | ruby phonology.rb
phonology : s/a, k/u, r/a, a, y/a, n/e
```

といった形で音素を取得することができる．

4.4. 0-1 整数計画問題としての文選択の定式化

ここまでの処理で，文候補に含まれるそれぞれの文に対してダイフォンが付与された．次に取り組む問題は「どの文を音素バランス文として採用するか」である．

最終的な目標は音声コーパスの構築であるので，選ばれた音素バランス文は声優に読み上げられ，録音されることになる．音声収録は一般的には声優ごとのキャリアやランクごとにワード (単語) や文字数あたりの価格が決まっているため，限られた予算内で音声コーパスを構築するためには

- 代表的なダイフォンをカバーし
- 総文字数，または総ワード数を最小化する

という二つの制約を満たすように文候補から音素バランス文として選ぶ必要がある．

声優統計コーパスの構築においては，

- 1 ワードの読みは最大 50 文字
- 料金はワードごとに発生する

という契約形態だったため，

- 選択されたバランス文の読みの総文字数を最大化する
- バランス文の文字数は最大 50 文字でなければならない
- 代表的なダイフォンをカバーしなければならない
- バランス文の総数は予算以下でなければならない

という制約にもとづき音素バランス文を選択する．

今回は文選択を 0-1 整数計画問題として定式化する．文候補 $\mathcal{S}$ に含まれる文字数 l_i の文 $s_i \in \mathcal{S}$ と，音素バランス文に含まれるべき N 個のダイフォンの集合 $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$ が与えられているとする．文 s_i にはダイフォン $\mathcal{D}_i = \{d_{(1,i)}, d_{(2,i)}, \cdots, d_{(N,i)}\}$ が付与されており，$d_{(n,i)}$ は 文 s_i に含まれるダイフォン n の数であるとする．また，全てのダイフォンについて，最低 M 回以上現れる必要があるというパラメータと，バランス文の総数を表す K というパラメータも導入する．ここで，文 s_i が音素バランス文として選択されるか否かを示すバイナリ変数 $x_i = \{0, 1\}$ を用いると，音素バランス文の選択問題は次のようになる．

$$\text{maximize} \quad \sum_i x_i l_i$$

$$\text{subject to} \quad \sum_i x_i d_{(j,i)} \geq m, \ \sum_i x_i = K, \ l_i \leq 50$$

ダイフォンをいくつまでカバーすべきかについては，文候補全てのダイフォンを数え上げ，頻出する上位 N 個を対象とするのが現実的であると思われる．また，最低登場回数 M については，得られた解と実際に支払うことができる予算との兼ね合いで増減する必要があるだろう．

今回は $N = 500$ すなわち上位 500 種までのダイフォンを対象とし，最低登場回数については $M = 3$ とした．

4.5. 文選択の求解

無事に 0-1 整数計画問題として定式化できたので，あとはソルバに任せるのみである．職業研究者であれば Gurobi[11] や Numerical Optimizer[12] などを用いるのであろうが，それらが利用できる環境にあればそもそもがこんなことをする必要はないだろう．

[9] http://julius.osdn.jp/
[10] gramtools/yomi2voca/yomi2voca.pl.in
[11] http://www.gurobi.com/
[12] https://www.msi.co.jp/nuopt/

無料で研究目的の利用が可能な整数計画問題ソルバとしては SCIP[13] や lp_solve[14]，GLPK[15]，Cbc[16] が存在している．いくつか試した限りでは，制約を満たしつつそこそこの精度の解を求めるまでの計算時間が速いことから SCIP を採用した．

4.6. 人手による修正・ヒューリスティクス

また，処理の概要としてはこれで十分であるが，実際にはさまざまな落とし穴がある．以下に注意すべき点や，用いたヒューリスティクスを記す．

読み付与の誤り 読み付与の誤りには二種類が存在している．一つは付与自体が失敗するケースである．これについては簡単なチェックで弾くことができる．問題は付与自体は失敗していないもの，正しい付与ではない場合である．先ほどの「佐倉綾音」や「水瀬いのり」の例がそれに該当するだろう．また，Neologd の中にもいくつか読みエントリが適切でないものが含まれているため，目視での確認は必要不可欠である．解の読みが誤っていた場合は，当該文を取り除き，ダイフォンの出現分布を数え上げ，制約を満たすよう手動で解を修正する必要がある．

数字や記号への対処 Neologd などを組み合わせることにより，数字や日付などについても MeCab は正しい読みを付与することが多い．しかし，読み上げ用コーパスとして考えた場合，それらの情報は必要なのだろうか．例えば，「大久保 瑠美（おおくぼ るみ、1989 年 9 月 27 日 - ）は、日本の女性声優」という文章について，括弧内を読み上げる必要はあるだろうか．こういった Wikipedia 特有の表記については，例えば括弧で囲まれた文字列そのものを除外する，といったルールベースで対応する必要がある．今回は，ひらがな，カタカナ，漢字，句読点，および長音符のみが含まれた文に限定することで対応した．

非文・難読文への対応 些細な事であるが，改行部分のみで文に分割した場合，箇条書きなどに正しく対応できない．よって，句点で終わる文のみを対象とした．また，それでもなお読み上げる文としてふさわしくない文が残るため，漢字のみ，カタカナと長音符のみで構成される文を除外した．

求解におけるヒューリスティクス SCIP が素晴らしいソルバであるとはいえ，しかし，Wikipedia 本文に含まれる文全てを対象に整数計画問題を解くのは現実的ではない．よって，そもそも文長の条件を制約に含めずにそもそも除外する，全文ではなくサンプリングした状態で解く，といったヒューリスティクスを用いた．

4.7. 音声収録

ソルバの奮闘の結果，声優に読み上げてもらうべき音素バランス文を得ることができた．あとは声優事務所や，事務所にコネクションを持つ代理店を通じて契約を締結し，読み上げてもらうだけである．今回我々は UWAN Pictures 社にキャスティングを依頼し，収録を行った．

ここで注意しなければならないのは，多くの声優事務所は収録した音声の無料配布を許諾しない，ということである．今回承知してくださった声優の皆様には感謝してもしきれない．

また，はじめて音声の収録を行ったが，そこでの気付きを記す．

- イントネーションを統一しなければならなかった・
- 息継ぎのタイミングを原稿レベルで統一すべきだった．
- 表現する感情の度合いをより緻密に検討すべきだった．

5. 結論

本稿では，いかにして独自の音声コーパスを構築するかについて述べた．今回は音素バランスしか考慮していないが，研究の目的によっては他にも考慮すべき項目，例えばトライフォンまで考慮した音素バランス文の選択や，感情表現の判定，演技の巧拙などが存在するだろう．また，ダイフォンの頻度という単純な制約のみではなく，より複雑な，それぞれの研究機関において秘伝のタレのように伝わる音声コーパスにおけるノウハウも存在しているだろう．それら無形の知見が，いつかなんらかの形で共有されることを切に願う．

本稿が，独自にコーパスを構築しようという人々にとって少しでも助けになれば幸いである．

Acknowledgement

キャスティングから音声収録までご協力いただいた UWAN Pictures 社，実際にご協力いただいた声優の土谷麻貴様，上村彩子様，藤東知夏様に熱くお礼申し上げます．

[13] http://scip.zib.de/
[14] http://lpsolve.sourceforge.net/5.5/
[15] https://www.gnu.org/software/glpk/
[16] https://projects.coin-or.org/Cbc

また，公開後に音声の編集ミス，ミス検出スクリプトを提供してくださった @shirayu 氏に感謝します．
最後に，日本声優統計学会の皆様，読者の皆様に感謝します．

References

[1] Kudo Taku. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.
[2] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.
[3] @y_benjo. 二次配布可能な音素バランス文と声優統計音声コーパスの構築. **声優統計**, 9, 2016.