1E5-2

音素バランスを考慮した読み上げ用フリー文章データベースの構築手法

Making a Database of Free Sentences to Read Considered Phoneme Balance

松永 寛之[*1] 橋本 直矢[*1] 佐々木 一磨[*1] 中臺 一博[*2] 尾形 哲也[*1]
Hiroyuki Matsunaga Naoya Hashimoto Kazuma Sasaki Kazuhiro Nakadai Tetsuya Ogata

[*1] 早稲田大学基幹理工学研究科 School of Fundamental Science and Engineering, Waseda University
[*2] ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan Co., Ltd.

This paper proposes a method to build a database for Audio-Visual Automatic Speech Recognition (AV-ASR) from copyright-free texts by considering phonemic balance. As a source of copyright-free texts, we utilized Aozora Bunko which is an open electronic library of Japanese novels on the Internet. We selected sentences from books in Aozora Bunko so that the number of phonemes can be balanced. After that, we asked 6 people to utter selected sentences and captured their voice sound and images for model training of AV-ASR. Experimental results showed the effectiveness of the proposed method..

1. はじめに

近年,機械学習分野で注目されている Deep Neural Network (DNN)は，音声認識精度の向上に有効であることが報告されている [5]. 我々は，DNN を利用して，視聴覚音声認識(Audio-Visual Automatic Speech Recognition, AV-ASR) の性能を向上することを目指しているが [6], 学習データが少なく, DNN の性能を最大限活かすことができないという問題があった. 本研究では，著作権フリー電子図書館「青空文庫」から AV-ASR 用の学習コーパスを構築する手法を提案する.

2. 関連研究

音声認識の研究に利用するデータベースを集める方法としては大きく分けて 2 通りの方法がある.

1つ目は定型の文章を音読して，収録したコーパスである. 代表的なものとして,「音声資源コンソーシアム」[1]で公開されている JNAS (新聞記事読み上げ音声コーパス)[2]が挙げられる. JNAS は約 60 時間程度の毎日新聞の記事と ATR 音素バランス文の読み上げ音声から構成されており，日本語大語彙連続音声認識の研究で一般的に用いられている. 定型文章の読み上げであるものの，話者によっては発話のゆれがあるため，ある程度の書き起こしを後処理として行っている. また，このコーパスの利用は研究目的に限られていること，音声・文章の著作権は保持されたままであるため，自由な編集，データの 2 次利用ができないことといった制約がある.

2つ目は自然な発話を収録し，後で音声データを書き起こしテキスト化する方法である. 日本語話し言葉コーパス（CSJ）[3]が代表的である. CSJ は日本語の講演などを録音して作られている. この方法は，自然な発話を収録するだけでよいので，大規模なデータを集めやすい. 一方で，書き起こし作業が大変，収録の雑音環境の制御が難しい，データ中の音素バランスは保障されないといった問題がある.

3. 文章データベースの構築

誰もが視聴覚音声認識研究に利用できるオープンなコーパスの構築を目指して，書き起こし作業が容易な定型文の読み上げベースで大規模なコーパスを構築する手法を提案する.

このような文章データベースを作るためには,

① 著作権フリーな文章の利用

② 音素バランスの考慮

の2つの条件が必要となる. ①はオープンなデータベースを作る上で欠かせない条件であり，②は音声認識で用いる音響モデル構築で鍵となる条件である. この2つの条件を満たすような文章データベースの作成方法を述べる.

3.1 著作権フリーな文の利用

著作権フリーな文章として，青空文庫[4]を利用した. 青空文庫は，ボランティアによってテキスト化された文学作品の電子図書館であり，インターネット上に公開されているため，だれでも無料でアクセスすることができる. 作品の著作権が消滅する作者の死後50年以降の作品が中心となっているが，作品数は約13，000点という膨大な量となっており，簡単に著作権フリーな文章を大規模に集めることができる.

3.2 音素バランスの考慮

音素とは，それぞれの言語においてそれ以上分けることのできない音の単位である. 音声認識では認識の最小単位として音素が用いられることが多い. 音素には様々な分け方があるが，一般的に日本語の音声認識によく利用される音素は以下の表 3.1 の39種類である[6].

表 3.1 日本語の音素一覧

母音:	/a/ /i/ /u/ /e/ /o/
	/a:/ /i:/ /u:/ /e:/ /o:/
子音:	/b/ /d/ /g/ /h/ /k/ /m/ /n/ /p/ /r/ /s/ /t/ /w/ /y/ /z/
	/ts/ /sh/ /by/ /ch/ /f/ /gy/ /hy/ /j/ /ky/ /my/ /ny/
	/py/ /ry/
その他:	/N/ /q/

音声認識の学習コーパスとして利用するためには，これらのすべての音素が含まれている文章データベースが好ましい. 文章中に各音素がいくつ含まれているのかを調べるために，自動的に日本語の文を音素列に変換するかな・音素変換器を作成した. これは，日本語の漢字かな交じり文を一旦カタカナに書き下し，対応する音素を割り当てるというものである. また，生成した音素列中に各音素がどれだけ含まれているのかをカウントする機能を付加した.

連絡先:松永寛之
早稲田大学基幹理工学研究科表現工学専攻
〒 169-8555 東京都新宿区大久保 3-4-1
TEL: 03-5286-2742
E-mail: matsuna4-10@akane.waseda.jp

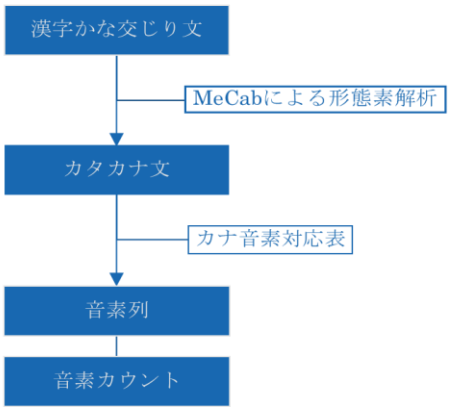

図 3.1 かな音素変換器の仕組み
例:サッカー選手
⇨ s a q k a: s e N sh u

このかな音素変換器を用い，1作品中にどれだけ音素が含まれているかを調べた．/by/,/gy/,/hy/,/my/,/ny/,/py/の6つの音素が含まれていないことが多く，単純に文章を選んでしまうとこれらの音素が極端に不足してしまうことがわかった．

これら6つの音素を含む文章を抽出し，意図的に足していくことで音素バランスがとれるよう図った．これら6つの音素は，以下の3種類の語句に多く含まれることがわかった．

①擬音語・擬態語
②外来語
③特定の漢字を使った単語

①の擬音語・擬態語は不足する音素6つを含む単語が多く見つかった．例えば，「ビュービュー」や「ぐにゃぐにゃ」，「ぴょんぴょん」，「ピューピュー」などである。擬音語・擬態語は小説系の作品中に多く含まれていることがわかった．

②の外来語も不足する音素6つを含む単語が多かった．「インタビュー」，「ヒューマン」，「ミュージック」，「ニュース」，「ポピュラー」など，多数の単語に含まれていた．

③については，例えば/gy/ならば，「業(ギョー)」は，産業・工業・業界のように多数の熟語中に含まれるため，「業」を探せば，多数の/gy/を見つけることができる．他の音素では，

/by/:病　　病気，病院
/hy/:評　　批評，評論，評判
/my/:妙　　奇妙，巧妙，微妙
/ny/:入　　先入，入選，入札

などが挙げられる．このような漢字を使った単語は評論文中に多く含まれていた．

3．3 構築した文章データベース

本稿では，手始めに，不足する音素がなく，またなるべく音素数のバランスが取れるように文章の選択を行った．結果として，合計50作品から358文を選択した．図3．2は選択した文章に対する各音素の含有数をヒストグラムで示したものである．比較的少数の文章で，バランスを考慮したコーパスの作成がシステマティックに実現できた．青空文庫の作品数の多さを考えれば，同様のプロセスを繰り返すことによって容易に大規模なコーパスを構築することが可能であると考える．

4. データベースの応用例，今後の展望

構築したデータベースの有用性を検証するため，視覚音声認識実験を行った．作成した文章データベースをタブレット端末

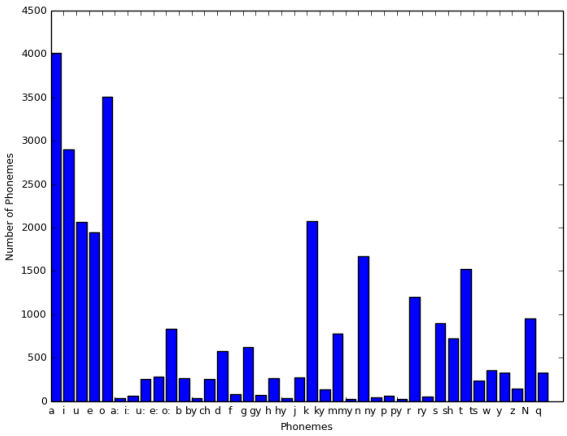

図 3.2 データベースの各音素含有数

上に表示し，読み上げ中の顔画像と音声を同時に収録するアプリを実装し，6人の視聴覚データ収集を行った．

本稿では1人分の収録データを，Convolutional Neural Network(CNN)を用い，隠れ層は 3 層で学習した．学習データに約 36000 枚，テストデータに約 3200 枚の画像を用いた．画像はサイズが 64x64，グレースケールのものを用いた．1 枚画像での認識率は約 21%であった．この結果は先行研究[6]と同等である．

5. 結論

本稿では，著作権フリーな文章を利用して，視聴覚音声認識に利用できるオープンなコーパスを作る手法を提案した．提案法を用いると，著作権フリーな文章として青空文庫を利用することよって，書き起こしコストの小さい読み上げベースで，かつ音素バランスを考慮して文章データベースが作成できることを示した．今後はデータ収集を継続して行い，大規模なデータの収集を行うと共に，構築したデータベースの有効性を視聴覚音声認識を通じて示したい．

謝辞 本研究は，文科省科研費基盤研究(A) (No.15H01710)の助成を受けた．

参考文献

[1]音声資源コンソーシアム(SRC)，2016.1.27
http://research.nii.ac.jp/src/
[2]JNAS (新聞記事読み上げコーパス)，2016.1.27
http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/
[3]CSJ(日本語話し言葉コーパス)，2016.1.27
http://pj.ninjal.ac.jp/corpus_center/csj/
[4]青空文庫，2016.2.2
http://www.aozora.gr.jp/
[5][Hinton 2012] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition”, IEEE Signal Processing Magazine, Vol. 29, pp.82-97, 2012
[6][Noda 2015] Kuniaki Noda, “Audio-visual Speech Recognition using deep learning”, Applied Intelligence, June 2015, Volume 42, Issue 4, pp722-737