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

1w120506－5　松永　寛之　　指導教員　尾形　哲也　教授

MATSUNAGA Hiroyuki　　Prof. OGATA Tetsuya

概要：　概要 : 本研究は，機械学習による音声認識を目的としたデータベースを収集するため，著作権フリーのテキストから読み上げ文章群を選定，データベース化する手法を提案するものである．具体的には，青空文庫に公開されている作品を自動的に音素化し，そのバランスを考慮した上で文章を抽出しデータベース化した．そして実際にこの文章群を用いて視聴覚データを収集し，深層学習モデルへの適用を行った．

キーワード：音声認識，文章データベース，音素

Key Words : Speech Recognition，Sentences Database，phoneme

1. はじめに

近年,機械学習の分野で注目されている Deep Neural Network(DNN)は，音声認識の精度の向上に有効であることが示されている．DNN の性能を最大限活かすには大量のデータが必要なのだが, 音声認識の様々な研究に対応できるデータが不足している．本研究では，音声認識を目的としたデータベースを収集するために必要な, フリーな読み上げ文章データベー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つ目は文章を音読しデータを収録するものである. 従来の日本語の音声認識の研究に利用されてきたものとしては，「音声資源コンソーシアム」[1]に公開されている JNAS(新聞記事読み上げ音声コーパス)[2]が代表的である．JNAS では，毎日新聞の記事と ATR 音素バランス文を読み上げ文章として利用している. 毎日新聞のテキストは大規模であり, ATR 音素バランス文は文字通り音素バランスが取れた文章であり, 音声認識に有用であった. このコーパスは研究利用するためには, 有償で公開されているが，音声データ，文章にはそれぞれ著作権が主張されており, 自由に編集することはできない.

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]などが代表的で，会議や講演などで大規模な音声データを集めやすい利点がある．しかし，音素バランスは保障されない.

３．文章データベースの構築

本研究では誰もが音声認識の研究に利用できるオープンなデータベースを作るために大規模でフリーな読み上げ文章データベースを構築する手法を提案する. 文章データベースを作るためには,

①著作権フリーな文の利用

②音素のバランスの考慮

の２つの条件が必要となる. この２つの条件をクリアする手法を述べる.

①著作権フリーな文の利用

著作権フリーな文章としてはいくつか考えられるが, 今回は青空文庫[4]を利用した. 青空文庫はインターネット上の図書館のようなものであり, 文学作品などがボランティアの方の手によってテキスト化され公開されていて, だれでも無料でアクセスし作品を読むことができる. 作品の著作権が切れるのは作者の死後５０年なので, 必然的に古い作品になるが, 著作権が消滅した作品だけで約１３０００点という膨大な量の作品が公開されており, 大規模な文章を集めるのに適している.

②音素バランスの考慮

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, 一般的によく利用される日本語の音素は以下の表 3.1 の３９種類である[5].

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/

文章データベースにはこれらのすべての音素が含まれていなければならない. 文章中にどれだけ音素が含まれているのか調べるために, 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 python で作成した. 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.

例：サッカー選手

⇒ s a q k a: s e N sh u

そして各音素の含有数を自動で調べるものを作り, 1 作品中にどれだけ音素が含まれているかを調べた. 数々の作品を検証した結果, /by//gy//hy//my//ny//py/の 6 つの音素が 1 作品中には含まれていないことも多く, 極端に不足するという傾向があることがわかった.

これら 6 つの音素を含む文章を抽出し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 つの音素は特定の漢字（病, 業, 百など）を使った単語, 擬音語・擬態語（にゃー, ぴょんなど), 外来語（インタビュー, ニュースなど）によく含まれていることがわかった.

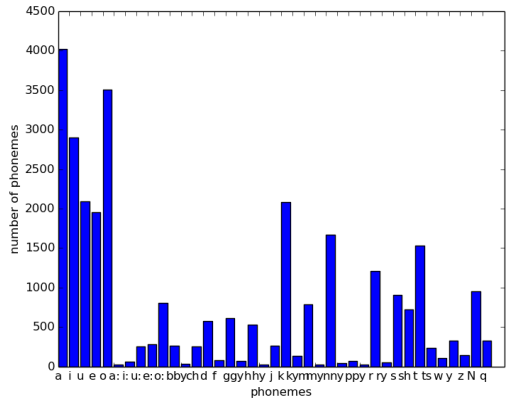

図 1　データベースの各音素含有数

①②の条件を満たすように文章を集めたところ, 図 1 のような音素数の文章データベースができた. 今回は不足する音素を補いながらも大規模な文章データベースを構築することを目指した. 文章データベースには合計５０作品から３５８文を使用した.

４．データベースの応用例，今後の展望

雑音に頑健な音声認識の 1 つとして, 通常の音声情報に読唇（リップリーディング）により得られる唇画像情報を付加する視聴覚音声認識というものがある. この研究を行うには人が話す音声と唇の動きが入った視聴覚データが必要である. 作成した文章データベースをタブレット端末に実装し, 倫理委員会の承認を得て, 視聴覚データ収集を行った. そしてこの集めたデータをリップリーディングの学習に利用した. 学習に利用することができることが示せたが, まだ有用な結果は得られておらず，今後の課題としたい.

５．結論・今後の展望

本研究では音声認識研究に利用できるオープンなデータベースを作るため, フリー読み上げ文章データベースを構築する手法を提案した. 青空文庫の著作権フリーな文章を使用し, 音素バランスを考慮しながら文章データベースを作成した.

今後はデータ収集を継続して行い, 集めたデータを用いて視聴覚音声認識の研究を行う. 大量のデータを集め, 学習させることで認識性能の向上を目指したい.

注：

*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 Kuniaki Noda, “Audio-visual Speech Recognition using deep learning”, Applied Intelligence, June 2015, Volume 42, Issue 4, pp722-737