研究報告

音声合成システムのための同形異音語の読み分け

梅村祥之，清水司

Japanese Homograph Disambiguation for Speech Synthesizers

Yoshiyuki Umemura, Tsukasa Shimizu

要　旨

文字情報を音声に変換するテキスト音声合成技術の研究開発が進んでいるが，漢字仮名混じり文を読み誤るという問題がある。読み誤りの実態を調べるため，大量の文章例から読み誤り率を評価するシステムを開発した。これを用いて，市販の音声合成システムの読み誤り率を測定し，読み誤り傾向を分析した結果，読み誤り原因の多くを同形異音語の読み誤りが占めていることが判明した。同形異音語とは，例えば「今日」が「きょう」と「こんにち」の2通りの読みを持つように，同一表記で複数の読みを持つ語である。同形異音語を文脈の情報から読み分ける技術として，決定リストによる方法が提案されている。この方法に改良を加え，高い読み分けの正解率を達成した。

キーワード　自然言語処理，音声合成，読み誤り，同形異音語，読み分け

Abstract

In recent years, text-to-speech systems have been advanced, but they sometimes make mistakes when reading Japanese sentences including the kanji characters. Therefore, we developed an estimating system using a misreading ratio.

We then measured the misreading ratio of a commercial text-to-speech system, and analyzed the tendency of the misreadings. As a result, we found that the main reason for the misreadings of determing in the reading of homographs. Japanese homographs sometimes stand for words that have several meanings when writing, for example, '今日' has two readings, 'kyoo' and 'kon-nichi'. Decision lists are used as a method to distinguish the homographs. We improved this method and achieved a high correct ratio.

Keywords　Natural Language Processing, Speech Synthesizer, Misreading, Homograph, Homograph Disambiguation

１．はじめに

カーナビゲーションシステムを初めとする種々の情報機器が自動車に搭載され，様々な情報通信サービスが始まりつつある。提供される情報には，交通情報，電子メール，新聞記事等があり，テキスト音声合成技術の重要性が増している。文章を入力して音声に変換するテキスト音声合成技術の研究開発が進んでいるが，品質面で，現在まだ，次のような問題がある。

(1) 漢字仮名混じり文を読み誤る
(2) 韻律 (アクセント，イントネーション，ポーズ) が不自然で，棒読みである
(3) 音声波形の音響的性質が機械音的で肉声感

に乏しい

本稿では，このうち (1) の仮名漢字混じり文の読み誤りを扱う。読み誤りの原因の中には，

● 同形異音語の判定誤り：

同形異音語とは，同じ表記で複数の読みを持つものである。例えば，「今日」は「きょう」と「こんにち」の２通りの読みを持つ同形異音語である。

● 固有名詞，時事単語の辞書未登録の問題

● 単語分割の誤り：

例えば，「米国産業界」において，米国 (べいこく) / 産業 (さんぎょう) / 界 (かい) という分割と，米 (こめ) / 国産 (こくさん) / 業界 (ぎょうかい) という分割があり得る。

などがある。このような原因が複合して読み誤りが生じるため，文単位の読み誤りが10%を越える音声合成システムが多い。

はじめに，現状の誤り率がどの程度であるかを評価するための，読み誤り評価システムについて述べ，それを用いて評価した結果を述べる。次に，読み誤りの改善方法について検討し，誤り原因中で上位を占める同形異音語の誤りに対して，これを改善するための読み分け方法について述べる。

２．音声合成システムの読み誤りの分析

２．１　読み誤り評価法の開発

音声合成システムから出力される発音情報から読み誤りを評価するシステムを構築する[1)]。そのためには，大量の文章を用いる必要があり，約20万文を収録した研究用文章データベースである EDR (Electronic Dictionary Research) コーパス[2)] を用いる。EDRコーパスに収録されている文には，**Table 1**に示すような付帯情報が付けられている。

付帯情報の中に「かな表記」があるので，読みの評価にこれを用いることができればよいが，次のような現象があり，そのまま用いることはできない。例えば「会社」が「○○会社」の文脈で「○○ガイシャ」となったり，「学校」の仮名表記「ガッコウ」が，発音の観点からは，「ガッコー」となるなど，仮名表記と，音声合成システムから出力される発音情報の間にいくつかの違いがある。

そこで，**Fig. 1**のように，EDRコーパスの仮名表記に，音声・音韻上の知見に基づく言語処理を行うことにより，音声合成システムがEDRコーパスの文を入力して作り出す発音情報中の読み誤りを評価できるようにする。言語処理は次の３項目である。

● EDRコーパスの仮名表記と音声合成システムの発音の不一致を，一義的に変換して差し支えない

Table 1　EDRコーパスの文例

構成要素番号	表記	かな表記	品詞	概念選択
1	会場	カイジョウ	名詞	3c0841
2	は	ハ	助詞	2621d5
3	熱気	ネッキ	名詞	102ab4
4	に	ニ	助詞	2621d5
5	包	ツツ	動詞	3ce654
6	ま	マ	語尾	2621cd
7	れ	レ	助動詞	2621c1
8	、	、	記号	2621d7

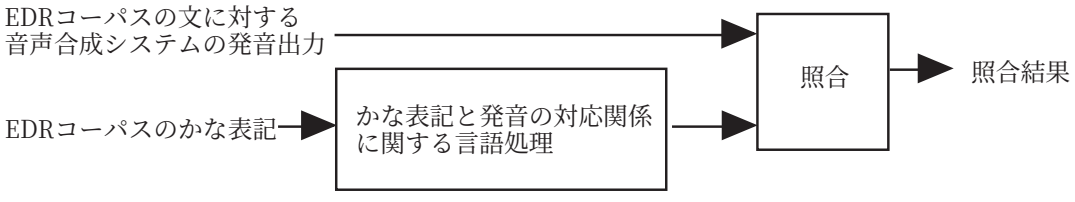

Fig. 1　読み誤り評価システムのブロック図

ものを変換する。

● 一義的に変換できないものについて，発音の候補をテーブル化し，照合段階でそのいずれかにマッチすればよいとして照合する。これに該当するものには，長音，連濁，英文字がある。

● 数字と助数詞に関し，判定不能扱いとし，照合段階で，それを避けるよう工夫する。

以下，各々の詳細を述べる。

２．１．１　一義的な変換

ここでは，一義的な変換によって仮名表記を発音記号に変換する例として，格助詞「を」を取り上げる。格助詞「を」の仮名表記はEDRコーパスで「ヲ」となっている。それに対して，市販の音声合成システムは,発音記号に「ヲ」を出力するものと「オ」を出力するものが存在する。「を」の発音は一義的に「オ」と考えてよいことから，あらかじめ,「ヲ」を「オ」に変換してから照合する。

２．１．２　発音候補テーブル

ここでは，発音候補テーブルによる処理のうち，長音を取り上げる。長音化の現象は，例えば「行進 (コウシン) 」の発音が「コーシン」のように，長母音の発音を生じる現象である[3)]。ところが，EDRコーパスの仮名表記は長音化を加味した表示になっていない。それに対して，音声合成システムは，長音化された発音を出力するため，照合の際にこれらは同じものであると見なす必要がある。しかし，仮名表記の情報のみから発音が長音化するかしないかを推定することは不可能である。なぜなら，例えば「子牛」と「格子」の仮名表記が「コウシ」であり同じであるのに対して，発音は「子牛」は「コウシ」,「格子」は「コーシ」というように異なり，最終的には語によって決まるからである。そこで，長音化の可能性のある語を計算機処理で洗い出し，音声合成システムから出力される発音と柔軟に照合する方式を採用した。

すなわち，仮名の繋がりから推定して長音化可能と考えられる部分をマークしておき，音声合成システムの出力が長音化していれば，EDRコーパスの仮名表記を長音化したものと照合し，長音化していなければ，EDRコーパスの仮名表記のまま長音化していない状態で照合する。仮名の繋がりから長音化の可能性を判定する処理は，次のように辞書引きによって行う。

仮名の繋がり：

"アア" "カア" "サア" "タア" "ナア" "ハア" …

それぞれについて，次のように長音化可能と見る。

長音化された仮名の繋がり：

"アー" "カー" "サー" "ター" "ナー" "ハー" …

２．１．３　照合

判定不能扱いの部分が存在する可能性があるため，**Fig. 2**のように照合する。すなわち，初め，文頭から形態素毎に照合して行き，判定不能形態素が現れたら，次に文末から文頭へ向かって形態素毎に照合して行く。この時点で再び判定不能形態素が現れたら，この文の処理を終了する。そのとき，前後の判定不能形態素に挟まれた判定可能形態素が存在すれば，準正解とする。 前後の判定不能形態素に挟まれた判定可能形態素が存在しなければ，完全正解とする。また，以上の照合の最中にミスマッチした形態素が始めて現れた時点で，その文の処理を終了し，不正解文としてカウントする。

以上の機能によりEDRコーパスの文例に対する音声合成システムの読み誤りを評価できるシステムを開発した。

２．２　読み誤り率算定と読み誤り傾向の分析

評価システムを，市販の音声合成システムに適用して読み誤りを評価した結果の一例を示す。音声合成システムから，発音と韻律の情報を持つ中間コードが出力されるので，その中の発音情報を用いる。以下で用いる量「完全正解数」「準正解数」「不正解数」は，前述の「照合」の節で述べ

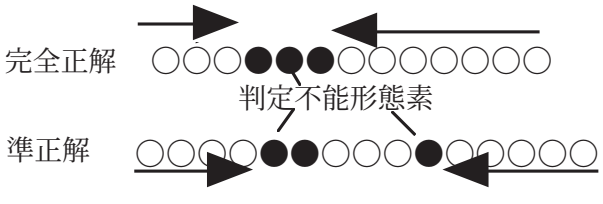

Fig. 2　照合方法

た量である。

評価した文の総数　＝　192,231文
完全正解数　＝　140,097文
準正解数　＝　12,831文
判定不能形態素に挟まれた判定可能な平均形態素数　＝　8.6
不正解数　＝　39,303文
1文当りの平均形態素数＝　24.4
1文当り正解率(準正解も正解に含める)=79.6%

Table 2に，本評価システムから得られた読み誤り例のうち，上位15の高頻度形態素を表示する。

以上，評価システムの結果から，

● 辞書の未登録あるいは登録誤りに関する大量の文例データが得られる。

● 誤り結果の分析から，誤りの種類が得られる

多く出現する誤りの種類は,「行った(イッタ／オコナッタ)」「今日(キョウ／コンニチ)」など同形異音語の読み分けに関するものである。また，頻度最大の助詞の「は」は，次の例のような，形態素の区切り誤りに起因すると思われる誤りである。

Table 2 読み誤り形態素の例

表記	品詞	仮名表記	出現頻度
は	助詞	ワ	2013
今年	名詞	コトシ	1384
私	名詞	ワタクシ	625
行	動詞	イ	566
ソフトウェア	名詞	ソフトウェア	487
その後	名詞	ソノゴ	485
他	名詞	タ	337
他の	連体詞	タノ	324
金	名詞	カネ	317
後	名詞	アト	316
言	動詞	イ	288
分	接尾語	ブン	279
後	名詞	ゴ	205
今日	名詞	キョウ	193
末	名詞	スエ	189

… 大統領選挙**は**いよいよ本番 …
　→ 大統領/選挙/**はい**/よい/よ/本番
… ダイトーリョーセンキョハイヨイヨホンバン …

さて，１番目の項目に関する改善のためには，この読み誤り評価システムから出力される読み誤り例を基に，未登録の語の登録と，登録誤りの修正を行えばよいため，それ以上踏み込まない。２番目の項目である同形異音語の誤りの改善のために，同形異音語の読み分け技術を開発した[4]。以下にその技術を述べる。

３．同形異音語の読み分け法

３．１　対象とする語の決定

日本語には数千語の同形異音語が存在する。総数を見積もるためには，辞書から，同一表記で読みの異なる語を抽出すればよい。形態素解析システム（文を単語に区切り品詞を求めるソフト）ChaSen 1.0[5]に付属の形態素辞書（国語辞典と同等な語彙数を持つ）から同形異音語の数を抽出すると，2,626語得られた。この中には，使用頻度のきわめて少ない特殊な語もかなり含まれている。そこで，約20万文を含むEDRコーパスから，頻度が少ない方の読みが10回以上出現する同形異音語に絞ったところ，362語になった。この中から，次の判断基準により，対象外の語を削除した。

● EDRコーパスの読み付与の問題で誤って同形異音語として抽出されたもの（例：「9月」の読みが「9がつ／くがつ」）

● 連濁によって複数の読みが生じたもの（連濁規則により，読みを求めればよいため[6]）

● EDRコーパスの単語区切りでは同形異音語となるが，ChaSenでは同形異音語とならないもの(例：EDRコーパスでは「運ぶ」が「運(はこ)/ぶ」，ChaSenでは「運ぶ」の1語)

● どちらの読みでもよいと思われるもの（例：「この間(このかん/このあいだ)」

その結果，204語に絞られた。

３．２　決定リストによる読み分け

決定リストを用いた読み分け方法[7, 8]の原理を述べる。基本的には，共起語を調べてそれから文脈の情報を得て判定する。例えば「今日（きょう

/ こんにち) 」の読み分けにおいて，次の例：

今日　の　天気　は　...

今日　の　日本　経済　は　...

のように，「今日」の近傍に「天気」があれば「きょう」と読み，「経済」があれば「こんにち」と読む。このように，「きょう」と共に現れる語を「きょう」の共起語という。実際は，共起語以外の情報も使うが，詳細は後の節で述べる。

ここで，いかにして共起語を得るかが問題になる。決定リストのアルゴリズムは，学習フェーズと判定フェーズからなり，学習フェーズにおいて，読みの与えられた大量の文例から次の計算で尤度比を求めて，読み分けの証拠能力の高さの指標としている。例えば「今日 (きょう) 」の共起語「天気」の尤度比の計算法を示す。「天気」が「きょう」の文例に N_{kyou} 回出現し，「こんにち」の文例に N_{kon} 回出現したとすると，尤度比は

$$\log_2 (N_{kyou} + 0.5) / (N_{kon} + 0.5)$$

として計算する。このようにして，学習用の文例に登場する全ての共起語の尤度比を計算し，それらを尤度比の絶対値の大きい順 (証拠能力の強い順) に並べてリストにする。これを決定リストといい，**Table 3** のような内容となる。リストの最後にデフォルトの尤度比を設定し，それ以下の尤度比の証拠は除去する。デフォルトの尤度比は，共起語に関係なく学習文例中に例えば「きょう」の文が1000文，「こんにち」の文が500文なら，$\log_2(1000.5/500.5)$として求める。

次に判定フェーズについて説明する。入力文を形態素解析して単語に区切ったものを用意し，決定リストの上位 (尤度比大) の証拠から順に，入力文に該当するものがあるかどうか調べて行き，合致するものが見つかった時点で，その証拠に対する判定結果を出力する。もし，何も見つからない場合には，デフォルトの判定結果を出力する。

Table 3 決定リストの内容の模式例

種類	内容	判定	頻度 (きょう)	頻度 (こんにち)	尤度比
	：				
共起語	経済	こんにち	8	150	−4.2
	：				
共起語	天気	きょう	100	15	2.7
	：				
デフォルト	―	きょう	1000	500	1.0

３．３　新しい決定リスト生成手法

３．３．１　決定リスト適用上の問題

今，「今日」を「きょう」と「こんにち」に読み分ける場合を考える。「きょう」の共起語，「こんにち」の共起語として容易に連想できる語がEDRコーパスを用いた学習データ中に何文含まれているかを調べてみる。例えば，**Table 4** に示すように，「きょう」から連想できる5語，「こんにち」から連想できる5語に関して，EDRコーパスに現れる頻度は十分でない。例えば，「休み」が共起語として得られないため，「今日休みます」という文の読み分けに失敗するであろう。したがって，学習データの量を増やせば，こういった共起語を獲得して，正解率が向上する可能性があると予想される。

３．３．２　学習データ増加法

「関連表現」を基に，文を機械的に抽出し，それを本来の語に置き換えて学習データとする。例

Table 4 「今日」の読み「きょう」と「こんにち」それぞれに関する共起語

共起語	きょう	こんにち
＜「こんにち」から連想＞		
経済	0	11
日本	3	24
状況	0	5
危機	0	2
低迷	0	0
＜「きょう」から連想＞		
試合	5	0
天気	3	0
運動会	2	0
会	2	1
休み	0	0

えば「今日」を「きょう」と読むための学習データを獲得するのに,

昨日，明日，あす，きのう，明後日，あさって

などの関連する表現を使い，生コーパスから関連表現を持つ文を機械的に抽出する。例えば,「明日」による抽出で,

「明日は仕事も休みなので、思う存分テニスを楽しむつもりです。」

という文が得られる。この文の「明日」を本来の「今日」に置き換える。置き換えられた文は,

「今日は仕事も休みなので、思う存分テニスを楽しむつもりです。」

となり,「今日 (きょう) 」の学習用の文として用いても問題ない文になっている。このような関連表現によって得られた文を学習データに追加してから，従来の方法で，決定リストを作成すると，例えば前述の「休み」が共起語に採用される可能性がある。一方,「今日」を「こんにち」と読むための学習データを増加するためには，関連表現として「最近の」を用いればよい。以上の方法で,EDRコーパスから関連表現によって学習データ量を増加させると，先に調べた共起語の出現頻度は**Table 5**のように増加する。「今日」に限らず，関連表現を見つけるには，基本的には，分類語彙表[9]や角川類語新辞典[10]などのシソーラス辞書で類語を調べる。

4．実験および結果

4．1　決定リストパラメータ

決定リスト作成にあたり，前述の共起語を含め,次の7種の証拠を使う。これらは，文献[8]から1つ省いたものである。

Table 5 関連表記による共起語の頻度増加

共起語	関連表現のみ		「今日」と関連表現	
	きょう	こんにち	きょう	こんにち
＜「こんにち」から連想＞				
経済	2	21	2	32
日本	10	15	13	39
状況	0	5	0	10
危機	0	1	0	3
低迷	0	1	0	1
＜「きょう」から連想＞				
試合	6	0	11	0
天気	5	0	8	0
運動会	2	0	4	0
会	10	4	12	5
休み	2	0	2	0

Table 6 複数種類の証拠を使った決定リストの内容の例

種類	内容	判定	頻度 (きょう)	頻度 (こんにち)	尤度比
	：				
共起語	経済	こんにち	8	150	−4.2
	：				
直後の1文字	は	きょう	100	15	2.7

読み分け対象の形態素（単語と考えてよい）の直前の1文字，直後の1文字，直前の品詞，直後の品詞，直前の形態素，直後の形態素，近傍の自立語(共起語)

なお，「近傍の自立語」における近傍の範囲とは，前後に自立語10個の範囲とする。学習フェーズにおいて，これら7種類の証拠に関する上記の対数尤度がそれぞれ計算され，対数尤度比の絶対値の大きい順に並べられる。したがって，上記の決定リストの内容としては，**Table 6**のように何種類かの証拠が混在する。

４．２　実験結果

3つの同形異音語，「今日(きょう，こんにち)」，「表(おもて，ひょう)」，「仏(ふつ，ほとけ)」について，EDRコーパスの文からの学習により決定リストを作成して読み分けを行い，正解率を測定し，さらに，前述の関連表現によって，文の数を増加させ，それを基に学習して決定リストを作成して読み分けを行い，正解率を測定する。なお，正解率算出にあたって10 fold cross validationを用いた。これは，学習時点と読み分け時点で，同じ文を使わないように文を分けて測定する方法である。結果を**Table 7**に示す。読み分けの難易度が高いと考えられている「今日」に関しては，関連表現の手法により読みの誤りが5.3%減少，関連表現が効果的に作用している「仏」に関しては，誤りが25%減少している。

５．まとめ

テキスト音声合成システムで漢字仮名混じり文の読み誤りが問題となっている。現状技術での誤り率を測定し，誤り傾向を分析するために，約20万文を含む研究用文章データベースであるEDRコーパスの文例から読み誤り率を算出し，誤り文例データを出力する読み誤り評価システムを開発した。これを用いて，市販の音声合成システムを評価し，誤り率を調べ，誤り傾向を分析した。その結果，文あたり2割程度の読み誤りを生じること，誤り原因として同形異音語による誤りが多いことがわかった。

そこで，同形異音語を読み分けるアルゴリズムを検討した。これまで提案されている決定リストを用いた読み分け法の正解率を向上させるには，学習データの増加が有効と考え，それを機械処理によって獲得する方法として，関連表現を用いる方法を提案した。この手法により，これまでの決定リストに比べ，難易度の高い同形異音語「今日(きょう，こんにち)」で，誤りを5.4%，「仏(ふつ，ほとけ)」で誤りを25%減少させた。

謝辞

本研究を行うにあたり，御協力いただいたトヨタ自動車(株)の遠藤徳和氏，青島滋樹氏に感謝致します。また，当所の原田義久氏に協力していただいた。

参考文献

1) 梅村祥之, 清水司, 原田義久 : "EDRコーパスを対象とした音声合成システムの誤読評価" , 言語処理学会第5回年次大会発表論文集, (1999)
2) 日本電子化辞書研究所 : "EDR電子化辞書仕様説明書" , (1995), 日本電子化辞書研究所
3) 日本語発音アクセント辞典, NHK編, (1985), 日本放送出版協会
4) 梅村祥之, 清水司 : "決定リストによる同形異音語の読

Table 7 決定リストと関連表現による読み分けの誤り率

語	読みと頻度	関連表現の文数	誤り%	誤り%（関連表現）	誤りの減少%
今日	こんにち 295, きょう 212	こんにち 371, きょう 190	18.7	17.8	5.3
表	ひょう 240, おもて 68	ひょう 404, おもて 182	19.2	10.6	21.4
仏	ふつ 166, ほとけ 12	ふつ 528, ほとけ 168	9.0	6.7	25.0

み分け" , 言語処理学会第4回年次大会発表論文集, (1998)
5) 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明 : "日本語形態素解析システム『茶筌』version 1.0 使用説明書" , 奈良先端科学技術大学院大学テクニカル・レポート, NAIST-IS-TR97007 (1995)
6) 佐藤大和 : "複合語におけるアクセント規則と連濁規則" , 講座日本語と日本語教育 第2巻 日本語の音声・音韻 (上) , 杉藤美代子編, (1989), 明治書院
7) Yarowsky, D. : "Decision lists for lexical ambiguity resolution:Application to Accent Restoration in Spanish and French", Proc. 32th Annu. Meet. Assoc. for Computational Linguistics, (1994)
8) 李航, 竹内純一 : "証拠の強さと信頼度を考慮した日本語同形異音語の読み分け" , 情報処理学会自然言語処理研究会資料97-119, (1997)
9) フロッピー版　分類語彙表, 国立国語研究所編, (1994), 秀英出版
10) 大野晋, 浜西正人, 角川類語新辞典CD-ROM版, (1989), 角川書店

(1999年9月28日原稿受付)

著者紹介

梅村祥之　　Yoshiyuki Umemura
生年：1957年。
所属：感性・心理研究室。
分野：自然言語処理, 音声情報処理に関する研究。
学会等：言語処理学会, 日本音響学会, 情報処理学会会員。

清水司　　Tsukasa Shimizu
生年：1970年。
所属：情報インターフェース研究室。
分野：自然言語処理，音声対話に関する研究。