外国語発音の自動評定と読み誤った単語の自動検出

Tobias Cincarek[1,2], *Rainer Gruhn*[1], *Christian Hacker*[2], *Elmar Nöth*[2] *and Satoshi Nakamura*[1]

[1] ATR 音声言語コミュニケーション研究所
[2] エアランゲン大学、ドイツ
{tobias.cincarek,rainer.gruhn,satoshi.nakamura}@atr.jp

はじめに

発音の自動評定とは、非母国語話者の音素及び単語の発音、または、文の発話が母国語話者と比べ、どの程度異なっているかを、自動的に推定することであり，外国語発音の学習における利用が期待されている。従来、発音の評定は、音素，文，文章のそれぞれのレベルで別々に行われていた[1, 2, 3]。本論文では、複数のレベルで発音を包括的に自動評定する方法を提案する。さらに、読み誤った単語を検出する方法についても述べる。

1. 発音の要素

文や文章を評価対象とすると、より多くの音素、単語を用いて評定を行うため、その話者の発音習熟度を高い信頼性で推定できる。さらに，発音評定の一つの基準となる流暢さを推定することができる。しかしながら，学習者にとって最も重要な要素である、どの単語を読み誤ったか、どの音素の発音が向上すべきか、という問題の発見、解決が難しいという欠点がある。そこで、本論文では文と単語の二つのレベルを対象にした評価を行った。

発音には、大きく分けると、三つの側面がある。具体的に述べてみると、（１）時間的な要素である発声率や単語間休止や音素及び単語の継続時間など、（２）他の韻律的な要素である文の音調や単語音節強勢など、（３）分節的な要素である母国語話者モデルを基準にした音韻的な類似性などがある。従来の研究[4]では、（２）の要素は、他の要素と比較すると、話者の発音能力の推定に関してあまり重要でないことが示されていたので、本研究でも考慮しないことにした。

2. 発音特徴抽出

前節の（１）と（３）の要素に対する定量的な評定を行うために，音声認識器で得られる分析値に基づいて特徴量を定義した。図１に発音特徴抽出の過程を示す．音響モデルの学習は WSJ コーパスで、音素列言語モデルと音素継続時間の分布の推定は TIMIT コーパスで行った。また音声認識エンジンとして HTK を用いた。

ある発声文に対して、音素レベルのアラインメントと、N-best 単語認識（単語ループ）を行い、得られた各音素の継続時間及びスコア（即ち音響モデルに対する尤度）と、認識した単語列とその該当する音素列から、音素列言語モデルと音素継続時間統計と共に、各単語と文全体の様々な発音特徴を求める（表１を参照）。

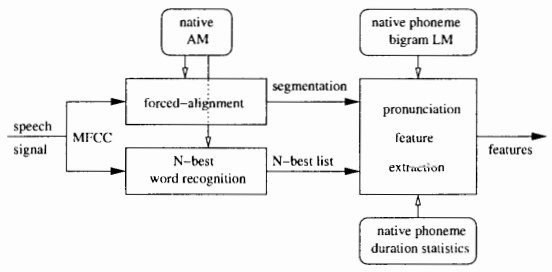

Figure 1. 発音特徴抽出

ID	特徴量	基本特徴の説明
1	音素尤度	発話に対する音響モデルの尤度
2	音素尤度比	アラインメントと認識仮説の尤度比
3	認識率	単語認識率或いは音素認識率
4	発声率	時間当り（秒）に発声する音素数
5	継続時間スコア	各音素の継続時間の尤度
6	音素列確率	認識音素列に対する言語モデルの尤度
7	音素継続時間比	実際の音素継続時間とその平均値の比
8	音素混同尤度比	分節認識信頼度（[5] を参照）
9	単語事後確率	N-best 仮説に基づく認識信頼度
10	発声率変動	単語ごとに変動する発声率
11	単語間無音長	言い淀みによって生じる単語間休止

Table 1. 発音自動評定に用いた特徴量

特徴量１～６は単語と文レベル両方、特徴量７～10 は単語レベルのみ、特徴量 11 は文レベルのみに対応する。特徴量２は [1] で提案された GOP スコアに基づいている。特徴量１～５は発音の評価に適していることが既に従来の研究 [1, 2] で示されている。特徴量１、２、５は音素レベルに対応するスコアであるが、各音素のスコアを累積することで、単語と文レベルのスコアとして用いた。これらは音素継続時間，音素数，発声率のそれぞれで正規化した。

3. 文と単語の評価

文の自動評定は、離散及び連続的な評価値付けが考えられる（図２を参照)。離散的な評価では、発音習熟度クラス毎に発音特徴のガウシアン分布を推定し、ガウシアン識別機で自動評定を行う。一方、発音特徴を線形変換することで、連続的な評価値を得られる。変換係数は線形回帰で求める。

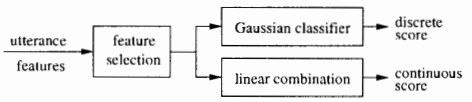

Figure 2. 発声文の自動評定

発音誤りの検出は単語毎に行う。しかし、単語のラベルを分析した結果（考察を参照）では、発音誤りの境界線が明白ではない、という事実が明らかになった。従って、自動検出において、その境界線をどう設定するかが重要である。

そこで、単語のより細かい区別を得るため、クラス「正しい」とクラス「発音誤り」の上に、クラス「不確定」を導入した。各単語は特徴ベクトルで表し、クラスごとにガウシアン分布を推定する。識別機において、ある単語の特徴ベクトルに対する各クラスの尤度を求め、もっとも尤度の高いクラスに割り当てる（図３を参照)。

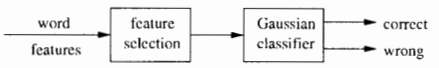

Figure 3. 読み誤った単語の自動検出

*Pronunciation scoring and extraction of mispronounced words for non-native speech.

なお、文の自動評定と発音誤りの検出において、定義した発音特徴の組合せを検定するために、いわゆる「floating search」([6] を参照) を適用した。その探索法は、逐次に特徴を加えながら識別機の性能を評価し、優れている特徴部分集合を出すものである。

4. データ

非母国語話者 96 人（うち大多数は、日本人、ドイツ人、フランス人、中国人、インドネシア人）から TIMIT の SX 文章 (48 文、約 400 語) の読み上げ音声を収録した。英語教師 15 人 (北米出身) は文毎に 1（最良）から 5（最悪）までの離散的な発音習熟度を示すラベルを付けた。その上で、読み誤った単語をマークした。この評価において、全ての話者と教師を四つのグループに分けた。ラベルの信頼性に関しては [7] を参照。最終的な文の評価値としてラベルの平均値を用いる。単語のラベルの分類に関して二つの方法を用いた。

- 分類A：教師二人以上にマークされた単語をクラス「発音誤り」に、残った単語をクラス「正しい」に統一する。
- 分類B：Aのクラス「正しい」を更に分け、一切マークされていない単語を新しいクラス「正しい」に、教師一人のみにマークされた単語をクラス「不確定」にまとめる。

実験において、三つのグループのデータを学習のために、一つのグループのデータを評価のために使用する。このようにして、最終的な実験を 4 交差検定で実施する。

5. 結果と考察

表 2 は文の自動評定の実験結果を示す。それによると、人間である教師にとって、発音習熟度に関して分節的な要素が一番大事である。教師に対する時間的要素の影響も大きい。六つの時間と分節的な発音特徴を線形に組み合わせることで、教師と同様な確度で、文の発音自動評定が可能である。

単独特徴・特徴組み合わせ	相関係数
特徴２：音素尤度比	0.48
特徴３：音素認識率	0.45
特徴５：継続時間スコア	0.45
特徴４：音素発声率	0.36
特徴１～６の線形変換	0.60

Table 2. 発声文の自動評定の結果：性能は教師の評価値と自動評定値との相関係数で表示。平均教師間相関は自動評定と同じく、0.60 である。

教師によって誤った発音であるとマークされた単語は，教師によって若干差異が見られた。 教師三人のラベルで単語を方法Aによって分類して、残った教師一人で評価を行った。その評価を四つの可能な組み合わせのために繰り返し、各混同行列を求める。表 3 はその交差検定の平均行列である。

教師	正しい	発音誤り
正しい	91.9	8.1
発音誤り	43.4	56.6

Table 3. 教師による単語の発音誤り検出率

正しい単語の８％は発音誤りとして、発音が誤った筈の単語の 43 %は正しい単語として判定された。後者の誤差は外国語学習者にとって好ましくなくても、発音の学習を損なわないと言えるだろう。一方、前者のような誤差は大きくなることにつれて、学習者に悪影響を与える。従って、発音誤りの自動検出は、前者の誤差が小さくなるように設計しなくてはならない。

表 4 は自動検出の結果を示す。発音誤りを検出する性能は高いが、正しい単語の 28 %も発音誤りと判定された。

そこで、分類法Bにって単語を三つのクラスに分類し、識別機を設計した。表 5 にそれに該当する判別結果がまとめてある。最終的に「不確定」の判別結果を「正しい」と見なし、教師一人のみにマークされた単語を発音誤りとして扱えば、表 6 が得られる。

自動検出	正しい	発音誤り
正しい	71.2	28.8
発音誤り	28.4	71.6

Table 4. 分類法Aの判別結果: 単語の発音誤りの自動検出の平均性能。

自動検出	正しい	不確定	発音誤り
正しい	71.0	15.1	13.9
不確定	45.2	19.7	35.0
発音誤り	22.8	18.8	58.4

Table 5. 分類法Bの判別結果：単語の発音誤りの自動検出率

このようにして、28 %であった誤差は 14 %までも減少する。それ一方、誤った単語の検出率は 43 %になる。教師の確度（表 3）と比較すると、有望な性能であると言えるだろう。

自動検出	正しい	発音誤り
正しい	86.1	13.9
発音誤り	57.3	42.7

Table 6. 分類Bの判別結果において、列「不確定」を「正しい」に、行「不確定」を「発音誤り」に統一した場合の自動検出率

なお、単語の発音誤りの検出において単独の特徴のみを用いた場合、特徴量８が最も優れており、特徴量１は二番目に優れていた。

6. 結論

本研究では、発声文の発音習熟度の自動評定と発音の誤った単語を検出するための実験を行った。文の自動評定の性能は、教師による評価確度と同様であり、平均相関係数は 0.60 であった。読み誤った単語の 43 %を自動で検出することができ、正しい単語に対する判別誤差は 14 %であった。

7. Acknowledgments

The authors would like to thank Naoto Iwahashi for proofreading of the Japanese draft.

This research was supported in part by the National Institute of Information and Communications Technology.

References

[1] S.M. Witt and S.J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108, 2000.

[2] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30:121–130, 2000.

[3] N. Minematsu. Yet another acoustic representation of speech sounds. In *Proceedings of ICASSP*, volume 1, pages 585–588, 2004.

[4] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *Proceedings of ICSLP*, 2000.

[5] S. Cox and S. Dasmahapatra. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(7):460–471, 2002.

[6] H. Niemann. *Klassifikation von Mustern, 2. überarbeitete Auflage im Internet.* http://www5.informatik.uni-erlangen.de/niemann/homeg.tht/homegli1.html, 2003.

[7] R. Gruhn, T. Cincarek, and S. Nakamura. A multi-accent non-native english database. In *Proceedings of Acoustical Society of Japan*, September 2004.