2006－SLP－60（9）
2006／2／4

韻律情報を用いた話し言葉音声認識における フィラー検出の改良

阿部 悠[†],　広瀬 啓吉[††],　峯松 信明[†],

[†] 東京大学大学院 新領域創成科学研究科,　[††] 東京大学大学院 情報理工学系研究科
Tel.: 03–5841–6767,　Fax.: 03–5841–6648
{yu-abe, hirose, mine}@gavo.t.u-tokyo.ac.jp

あらまし　本稿では、話し言葉特有の現象である言い淀み (フィラー) を、より高精度に検出する方法を提案する。韻律を音声認識に利用する従来の研究では、大量のデータが得やすい朗読調音声を対象としてきた。しかし話し言葉は、データの用意が困難な上、休止や言い直しなど不規則な現象も頻繁に発生し、認識率が低い。これらの現象は韻律的に通常の発話部分とは異なる特徴を持つ。特にフィラーは比較的安定した韻律的特徴を有する。そこで我々は話し言葉音声認識への韻律利用の試みの一例として、フィラーに着目し、既存の大語彙連続音声認識システム (Julius) に、韻律的に判断したフィラーらしさを算出するモジュールを実装し、認識実験を行った。その結果、フィラー検出の向上を実現した。

キーワード　話し言葉音声認識、フィラー検出、韻律的特徴、機械学習、Julius

Improvement of Filler Detection Using Prosodic Features in Spontaneous Speech Recognition

Yu Abe[†],　Keikichi Hirose[††]　and　Nobuaki Minematsu[†]

[†] Graduate School of Frontier Sciences, University of Tokyo,
[††] Graduate School of Information Science and Technology, University of Tokyo
Tel.: 03–5841–6767,　Fax.: 03–5841–6648
{yu-abe, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract　This paper describes a new scheme of detecting fillers which are frequently included in spontaneous speech. So far, speech recognition researches using prosodic features have focused on text-reading style speech. In such cases, large amount of data are usually obtainable. However, the situation of spontaneous speech is different. Besides, spontaneous speech may include a number of irregularities, such as hesitations, re-statements, and so on, which may largely degrade speech recognition performance. These parts show prosodic features different from other parts of normal utterance. Especially, fillers have relatively stable prosodic features. As an example of attempt to use prosodic features for spontaneous speech recognition, we focused on fillers, and implemented the prosodic module which calculates filler likelihood in an existing large vocabulary continuous speech recognition system(Julius). We conducted speech recognition experiments, and realized the improvement of filler detection.

Key words　spontaneous speech recognition, filler detection, prosodic features, machine learning, Julius

1　はじめに

韻律的特徴は様々な要因で多様に変化する。このため、これまでの音声認識の研究では、韻律はノイズとして積極的に排除される傾向にあった。しかし、我々人間が音声を認識する過程では、韻律が非常に重要な役割を果たしていることは明らかである。

このような観点から、近年、韻律を音声認識に積極的に利用しようとする研究が行われるようになってきた。大語彙連続音声認識に韻律を利用し、ある程度の認識率の向上を実現した先行研究としては [1–4] などがある。しかしこれらの研究に共通して言えることは、朗読調音声を対象としていることである。朗読調音声の場合、音響モデル、言語モデルの学習に必要な大量のデータが比較的入手しやすい。そのため、韻律的特徴を利用しなくてもそれなりに高い認識率が得られ、認識過程全体に対する韻律利用の効果がはっきりしない。

しかし、話し言葉音声となると、大量のデータを用意することは難しい上に、話し言葉は言い淀み、休止、言い直し、咳払いなど、不規則な現象をたくさん含んでいる。これらは、音声認識の精度を著しく低下させる原因となる。しかしこれらの現象は、他の通常の発話部分とは違った韻律的特徴を見せる。したがって、それらの部分は $F0$ パターンや、パワー、音素持続時間等を調べることで検出できる可能性がある。そしてその結果得られた情報は最終的な認識率の向上に利用できるかもしれない。例えば、これらの現象を検出できれば、入力音声からその箇所を除去した上で認識過程を進めるといったことが考えられる。したがって話し言葉音声認識において、韻律の利用が朗読調音声認識よりもさらに有効であると考えられる。本研究ではその中でも比較的韻律的特徴がはっきりしているフィラー [5] に焦点をあてる。特にフィラーは人間の対話で positive な役割を果たすことも指摘されており [6]、フィラー検出により、対話システムにおける対話戦略への応用なども期待される。

本論文は、2 章で提案手法の概要について、3 章で実験に用いた音声データについて、4 章で韻律モジュールについて、5 章で認識実験とその結果について、そして 6 章でまとめを述べる。

2　フィラー検出の枠組み

本研究では音声認識エンジンとして Julius を利用する。Julius はオープンソースソフトウェアであり、提案手法を実装し、エンジンに組み込みやすいという利点がある。Julius は 2 パス探索を行い、第 1 (前向き) パスで簡易なモデルにより単語候補をしぼったうえで、第 2 (後向き) パスで高精度なモデルを用いて再探索、再評価を行う [7]。

この Julius を用いた提案手法の枠組みを図 1 に示す。韻律モジュールは、形態素のフィラーである確率スコアを計算する。この確率スコアを「フィラー度」と名付ける。韻律モジュールは全ての形態素に対してフィラー度を算出することが可能であるが、現段階では第 2 パスでフィラーと仮説が立てられた形態素についてのみ、フィラー度を算出する。韻律モジュールについては詳細を 4 章に述べる。算出されたフィラー度がしきい値を越えた場合、ボーナスとして言語スコアにある一定値を加算する。このボーナスを以降「韻律スコア」と呼ぶ。加算する韻律スコアをフィラー度に応じた形で変化させるか、一定値とするかは検討事項であるが、本実験ではフィラー度がしきい値 0.5 を越えた時に、韻律スコアを一定値 5 とした。詳細は 5 章に述べる。

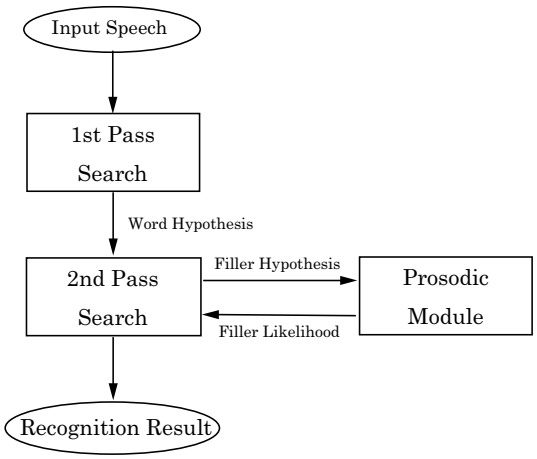

図 1.　提案手法の概要

3　音声データ

用意した音声データは 100 発声である。話者は男性 7 名、女性 6 名であり、各発声に 1 つ以上のフィラーが含まれている。使用したコーパスは日本語話し言葉コーパス (CSJ) である [8]。もともとのデータの形態は 1 つの学術講演が 1 つのファイルに収録されている。したがって、これらのファイルから、フィラーを 1 つ以上含む文を 100 発声切り出した。切り出す際に、言い直しや咳払いを含まないものを選択した。

CSJ のデータ全体で 160 種類のフィラーが出現するが、上記 100 発声の中には 17 種類 389 個のフィラーが含まれている。100 発声中に出現するフィラーとその出現頻度の例を表 1 に示す。

表 1.　100 発声中に出現するフィラーとその出現頻度の例

フィラー	出現回数
えー	185
え	82
そのー	16
ま	14
まー	13
えっと	12
あの	11

4　韻律モジュール

4.1　構成

韻律モジュールは機械学習を用いて構成する。本研究ではニューラルネットを採用した。学習の単位は CSJ の定義する短単位という形態素とした。ネットの構成は、中間層 3 層を含む 5 層構成となっており、中間層はそれぞれ 20 ユニットである。この形はいくつか行った予備実験を通して決定した。入力層は 10 ユニット、出力層は 1 ユニットである。入力層の各ユニットは表 2 に示す 10 種類の各韻律情報に対応している。この表で、特に何も表記がない項目は当該形態素についての項目である。また、$F0$ はすべて対数スケールである。出力層は 0 から 1 のフィラー度を出力する。

表 2.　機械学習に用いた韻律情報

韻律情報
音素数
$F0$ の最大・最小の差
$F0$ を一次直線で近似した時の傾き
(母音平均 $F0$)÷(全母音平均 $F0$)
最終母音と後続形態素の最初の母音の $F0$ の差
先行する振幅の弱い区間の長さ
後続する振幅の弱い区間の長さ
振幅を一次直線で近似した時の傾き
母音の振幅の平均
(最終母音の dur)÷(全音素の平均 dur)

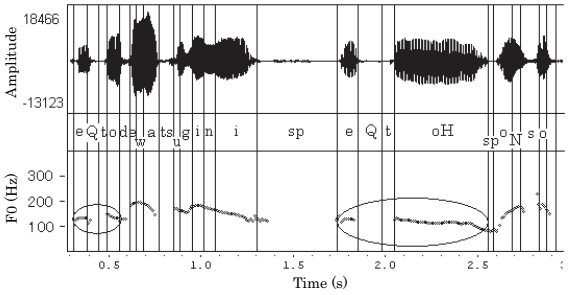

図 2.　フィラーの振幅と F0 パターンの出現例

図 2 はフィラーの振幅と $F0$ パターンの出現例を示したものである。「えっとでは次に、えっとー音素 · · ·」という男性の発声の一部分であり、$F0$ パターンの丸で囲まれた部分はフィラー箇所に対応する。この例も含めて、フィラーは、その $F0$ パターンはしばしば他の部分と比較して低く、かつ一定であることや、前後に無音区間があること、最終母音が引き延ばされることなどの韻律的特徴を持つことが多い。これは、フィラーの韻律的特徴を調査した先行研究 [5] とも一致する。表 2 であげた韻律情報はこれらを考慮して考えたものである。

4.2　ニューラルネット構築

4.1 の条件でニューラルネットを作成した。まず全ての発声について、Julian によるセグメンテーション、$F0$ の抽出、10ms ごとの RMS の算出を行った。そして表 2 にあげた韻律情報を算出した。ここで 12 発声についてはセグメンテーション、$F0$

の抽出のエラーなどにより、韻律情報を得られなかったので除外した。韻律情報が得られた 88 発声 (男性 6 名、女性 6 名) を、各話者 1 発声ずつの計 12 発声を評価用に残し、残りの 76 発声 (男性 4 名、女性 4 名) をニューラルネットの学習用とした。学習用 76 発声は計 2846 形態素からなり、うち 306 形態素がフィラーである。一方評価用 12 発声は計 420 形態素からなり、うち 39 形態素がフィラーである。図 3 はニューラルネットの学習回数と squared error の関係を示したものである。この図から、20 回程度の学習で十分収束していることがわかる。学習回数が増えすぎると過学習の影響によりニューラルネットの精度が落ちるという予備実験の結果も考慮して、本実験では学習回数を 50 回とした。

図 3. 学習回数と squared error の関係

4.3 性能評価

4.3.1 評価音声に対する性能

作成したニューラルネットを、評価用の 12 発声で評価した結果を表 3 に示す。この表は、出力層が 0.5 以上を出力したときに「ニューラルネットがフィラーであると判定した」と定義したときの、各発声の各形態素について、ニューラルネットがフィラーか否かを正しく判定できたかを集計したものである。見方は例えば男声 1 なら以下のようである。発声した一文には 48 個の形態素が含まれ、そのうち 5 個の形態素がフィラーである。作成したニューラルネットにより、48 個のうち 43 個を正しく判定でき、5 個のフィラーのうち 3 個を正しく判定できた (すなわち検出できた)。ちなみに全ての発声はテキストオープンであるが、表の太字の発話は話者オープンでもある。全体として 39 個のフィラーのうち、29 個を検出できたことが分かる。

表 3. ニューラルネットの性能評価

音声	フィラー	全形態素
男声 1	**3/5(60%)**	**43/48(90%)**
男声 2	**4/4 (100%)**	**37/37 (100%)**
男声 3	1/2(50%)	21/23(91%)
男声 4	2/3(67%)	30/31(97%)
男声 5	5/6(83%)	38/41(93%)
男声 6	1/2(50%)	21/23(91%)
女声 1	2/3 (67%)	23/25 (92%)
女声 2	2/3 (67%)	38/40 (95%)
女声 3	**3/3 (67%)**	**30/31 (97%)**
女声 4	1/3 (33%)	29/31 (94%)
女声 5	4/4 (100%)	58/59 (98%)
女声 6	**1/1 (100%)**	**29/31 (94%)**
計	29/39 (74%)	397/420 (95%)

表 4. 評価音声中の 39 個のフィラーの Julius による認識実験の結果

決定木	Julius	Julius 1st pass	個数
×	×	×	0
×	×	○	0
×	○	○	10
○	×	×	0
○	×	○	6
○	○	○	23

4.3.2 Julius への実装に向けた考察

韻律モジュールは Julius の第 2 パスでフィラーという仮説が立てられた形態素についてのみ、フィラー度を算出する。したがって、第 1 パスで認識できなかったフィラーに対しては、韻律モジュールは精度がどれだけよくても機能しない。そこで評価発声中に含まれるフィラーが Julius で認識させたときに、仮説に含まれているかは重要である。それを調べたものを表 4 に示す。なお、認識実験の際、音響モデル、言語モデルは CSJ が提供するもの [9] を利用した。表 4 の見方は、例えば、評価音声中に含まれる 39 個のフィラーのうちの 10 個は、Julius による最終的な認識結果で認識できたが、韻律モジュールはフィラーであると検出でき

なかった、ということである。この表によると、39個の全てのフィラーが第1パスの仮説の中に含まれていたことがわかる。また、本モジュールが最も効果を発揮するであろう状況、すなわち「Juliusの最終認識結果ではフィラーと認識されていないが、第1パスの仮説には残っている、なおかつニューラルネットではフィラーであると判定された」という状況は6個あった。逆に、悪影響を及ぼす可能性がある状況は、「Juliusの最終認識結果でフィラーと認識されていたが、ニューラルネットでは検出できなかった」であるが、これに該当するフィラーは10個あった。

このことを考慮し、5章で述べる認識実験では、韻律モジュールの結果、フィラーである確率が低い時にペナルティを与えることはせず、高い時にボーナスを与えるのみにすることが有効であると考えられる。

5　認識実験

5.1　実験条件

4章で述べた韻律モジュールをJulius ver.3.4.2に組み込み、用意した100発声を用いて認識実験を行った。音響モデル、言語モデルは4.3.2の時に用いたものと同じである。また、4.3.2の考察もあり、フィラー度が0.5を越えたときに韻律スコアを一定値5とし、ボーナスとして言語スコアに加算する一方、フィラー度が0.5を下回ったときはペナルティを与えることはせず、何も処理を施さなかった。

5.2　実験結果

韻律モジュール導入により、認識結果が改善された例を示す。例1は認識できなかったフィラーが認識できるようになったものである。

例1
正解：･･･仮説 が え 支持 さ れ ま し た 。
baseline：･･･仮説 が 認識 さ れ ま し た 。
proposed：･･･仮説 が え し さ れ ま し た 。

またフィラーでない形態素についても、韻律モジュール導入により例2のように改善された。

例2
正解：･･･え こちら が えー 鼻 の ある･･･
baseline：･･･え こちら側 えー 鼻 の ある･･･
proposed：･･･え こちら が えー 鼻 の ある･･･

100発声によるこれらの認識実験の結果をまとめたものが表5である。この表より、例1のように認識できなかったフィラーができるようになったものが7個、逆に本モジュールの悪影響により、認識できていたフィラーが認識できなくなったものは全くなかった。フィラーでないものについては、例2のような改善が4個ある一方、認識結果が悪くなったものも3個あることがわかる。ちなみに100発声中に含まれる389個のフィラーうち、baseline Juliusが認識できたものは349個、proposed Juliusでは356個である。

表5.　100発声を用いた実験結果

(baseline → proposed)	filler	Non-filler
誤 → 正	7	4
正 → 誤	0	3

5.3　考察

例1は表3における、ニューラルネットの評価用の12発声のうちの女声5の発声の結果である。表3より、女声5に含まれる4つのフィラーはニューラルネットにより全て検出できたことがわかるが、そのうちの1つが例1のフィラー「え」であり、これはbaselineのJuliusでは認識できなかったが第1パスの仮説には残っていたものである(すなわち表4における6個のうちの1つであった)。それが、認識過程においてもニューラルネットによりフィラー度が0.79という値が算出され、ボーナスを加算されたことで、「･･･仮説が認識されました。」という仮説をスコアで上回った。これは韻律モジュールが上手く機能していた例であると言える。

また、もともと認識できていたフィラーが認識できなくなったものは1つもなかった。これは、韻律モジュールの結果に対してボーナスのみ与え、ペナルティを与えなかったことによると考えられる。

一方、韻律モジュールの悪影響として、フィラーでない形態素で、認識結果が悪くなったものが3個ある。これらを減らす最も単純な方法は、ボーナスとして与える韻律スコアを小さくすることであるが、その結果、表5の7個という個数も当然減ってしまう。したがってあまり望ましい対策方法ではない。この悪影響の原因は、処理時間的、使用メモリ量的な側面からJuliusがかけている近似と大きく関係する。Juliusは、仮説の各長さごとに、展開された仮説数がしきい値を越えた場合に、それより短い仮説を展開しないようにするという、「仮説エンベロープの幅の制限」を設けている。この近似のため、正解仮説がスタック内に残っているにも関わらず、第2パスで展開されずに消滅していた。仮説エンベロープ幅はdefaultでは30であるが、この値を非常に大きく、すなわち全探索に近い条件に設定し、3発声を再度認識させたところ、この悪影響が解消されることが確認された。

6 まとめ

話し言葉音声認識において、フィラーを検出する新たな枠組みを提案した。韻律は様々な要因によって多様に変化する現象でありながら、それを音声認識に利用することで、精度を向上させることを実現した。

今後は、データサイズを数倍に拡大する。フィラーはその様子が話者により大きく異なることが知られている。したがって、提案手法を、多数の話者による様々な発話スタイルにも対処できる手法へと発展させていくことを目指す。

参考文献

[1] Hirose, K. and Iwano, K.: "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," Proc. IEEE ICASSP, pp.1763-1766, 2000.

[2] Lee, S., Hirose, K. and Minematsu, N.: "Incorporation of prosodic module for large vocabulary continuous speech recognition," Proc. ISCA Tutorial and Research Workshop on: Prosody in Speech Recognition and Understanding, pp.97-101, 2001.

[3] Hirose, K., Minematsu, N. and Terao, M.: "Statistical language modeling with prosodic boundaries and its use for continuous speech recognition," Proc. ICSLP, pp.937-940, 2002.

[4] Hirose, K. and Minematsu, N.: "Use of prosodic features for speech recognition," Proc. ICSLP, pp.1445-1448, 2004.

[5] Quimbo, F. C. M., Kawahara, T. and Doshita, S.: "Prosodic analysis of fillers and self-repair in Japanese speech," Proc. ICSLP, pp.3313-3316, 1998.

[6] Watanabe, M., Hirose, K., Den, Y. and Minematsu, N.: "Filled Pauses as Cues to the Complexity of Following Phrases," Proc. InterSpeech, pp.37-40, 2005.

[7] Lee, A., Kawahara, T.,and Shikano, K.: "Julius - an open source real-time large vocabulary recognition engine," Proc. EUROSPEECH, pp.1691-1694, 2001.

[8] Maekawa, K.: "Corpus of Spontaneous Japanese: Its design and evaluation," Proc. ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition, pp.7-12, 2003.
http://www2.kokken.go.jp/ csj/public/ index.html

[9] Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: "Benchmark test for speech recognition using the Corpus of Spontaneous Speech," Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.135-138, 2003.