E-024

話し言葉音声中のフィラー検出精度と音声認識率の関連性
Relationship between Performance of Filled Pause Detection and Speech Recognition in Spontaneous Speech

宗宮 充宏† 西崎 博光‡ 関口 芳廣‡
Mitsuhiro Somiya Hiromitsu Nishizaki Yoshihiro Sekiguchi

1. はじめに

本稿では，講義・講演音声中に含まれるフィラー[1]検出精度と音声認識率の関連性について述べる．

講義や講演音声にはフィラーが含まれており，これは話し言葉特有の大きな特徴である．例えば，フィラーは個人の話し方に大きな影響を与えており，話の聴きやすさ，理解しやすさに大きな影響を与えることが我々の研究で明らかになっている [1]．

そこで，話し方改善を目的としたフィラーの使い方自動評価についての研究を行っている [2]．この研究ではフィラーの検出精度が重要となることから，まず，フィラー検出精度の改善を図る試みを行っている．その過程において，フィラーの検出率とフィラー以外の単語も含めた全発話の音声認識率との間に高い関連性があることが分かった．

先行研究において，フィラー検出に関しての報告はあるが [3][4]，検出性能とフィラー以外の発話も含めた音声認識率との関連性について詳しい調査はされていない．本稿では，より詳細にフィラー検出率と全発話の音声認識率との関連性について分析を行った．

分析の結果，フィラー検出精度と全発話の音声認識率とは高い相関がある以外に，フィラーが発話のどの位置に出現するかによってフィラー検出精度に差がでること，この出現位置が全体の音声認識率に影響を及ぼしていること等が明らかになった．

これらの分析によって，話し言葉音声認識の認識率を改善をするための一つの可能性を見出したので，それについて報告する．

2. フィラーの分類と検出方法

本研究では，実験対象として大学の講義音声と日本語話し言葉コーパス (CSJ) を用いた．

2.1 フィラーの分類

発音が似ているフィラー同士[2]を一つのグループとした．そして，書きこし文に存在している 82 種類のフィラーを 10 グループに分類した．フィラーの検出は，正解文中にあるフィラーと全く一致していなくても，同じグループのもので認識されていれば正解とした．

2.2 フィラー検出方法

大語彙連続音声認識エンジン Julius を用いてフィラー検出を行った．フィラーとして認識された単語をそのまま利用することでフィラー検出とする．使用した音響モデルは CSJ の 797 講演から学習したトライフォンモデルである．また，言語モデルも CSJ から学習した 20k の単語トライグラムである．

2.3 フィラー検出の評価

フィラー検出の評価の方法は，精度 (*precision*)，再現率 (*recall*)，F 値 ($F-measure$) を用いる．フィラー検出の性能は F 値によって表される．

3. フィラー検出実験

講義音声は 28 講義 (話者 7 人)，CSJ ではオープンテストデータ[3]10 講演 (話者 10 人) とクローズドテストデータ[4]10 講演 (話者 10 人) を用いて[5]フィラー検出の実験を行った．音声の時間は，大学講義が 1 講義 90 分程度 (一部約 30 分の講義あり)，CSJ が 1 講演 12 分程度である．

フィラー検出率・フィラー出現位置の割合・音声認識率を表 1，それらの相関を表 2 に示す．表 1 における，「Head」,「Middle」,「Isolate」は正解文中のフィラーの出現位置の割合である．それぞれ，文頭に現れたフィラー，文中に現れたフィラー，単独で存在しているフィラー (音声中でフィラーしか発話されていない) を示している．また，「Corr.」は単語正解率，「Acc.」は単語正解精度を示している．ここでのフィラーの出現位置は，我々がこれまでの研究 [1] で着目してきたフィラーの特徴の一つである．

講義音声においては，同一講師による講義が複数個存在している．そこで，28 講義全てを用いた場合と，1 人 1 講義として 7 講義を用いた場合の結果を出した．それらを比較してみると，値に若干の差はあるが，傾向が大きく変化することはなかった．

表 1 からどの音声のフィラー検出率も，精度より再現率の方が高くなっている．つまり，多くのフィラーは正しく検出されているが，誤検出も多い結果となっている．そして，フィラー検出率の F 値が大きい音声ほど，音声認識率 (Corr., Acc.) が良くなっていることも確認できる．これについては，音声認識率を求める対象にフィラーも含まれているため，フィラーの検出が良くできていれば音声認識率が良くなるのは当然かもしれない．しかし，ほとんどの音声で全形態素中のフィラーの割合が 10 %未満である．従って，フィラーの認識率が音声全体の認識率に与える影響は小さいと言える．

表 2 から講義音声ではフィラー検出率の F 値と音声認識率の間には正の相関があることが分かる．特に，Corr. と Acc. の双方で非常に高い相関が現れた．また，オー

†山梨大学大学院医学工学総合教育部，University of Yamanashi
‡山梨大学大学院医学工学総合研究部，University of Yamanashi

1) 「えーと」や「あのー」などを示す．

2) 例えば，「えっと」と「えーと」など．

3) CSJ 付属の「test-set 1」を利用した．

4) 正解文中に含まれるフィラー数がオープンテストデータと同じ程度の音声を集めた．

5) 講義音声は無音区間を基準とした文献 [5] の方法，CSJ では 0.2 秒の無音を基準とした方法で機械的に分割された音声を扱った．

表 1: フィラー検出率・フィラー出現位置の割合・音声認識率

試料	フィラー検出率			フィラー出現位置の割合 [%]			音声認識率 [%]	
	Precision	Recall	F-measure	Head	Middle	Isolate	Corr.	Acc.
CSJ(クローズド)	0.68	0.84	0.75	42.15	50.29	7.56	76.28	71.75
CSJ(オープン)	0.60	0.75	0.67	48.95	43.43	7.62	66.44	58.79
講義音声 (28 講義)	0.51	0.85	0.64	49.55	38.82	11.63	55.55	46.02
講義音声 (7 講義)	0.52	0.83	0.64	48.38	39.67	11.95	54.83	45.07

プンテストデータは，Corr. よりも Acc. との相関が高い結果となった．

一方で，クローズドテストデータは無相関に近い結果であった．クローズドテストデータはどれも音声認識率が高いため，オープンテストデータや講義音声と比べて Corr. と Acc. の分散が低いために相関が現れにくいと考えられる．ただし，オープンテストデータ，クローズドテストデータではフィラーの数が非常に少ない (10 個未満) ものが一講演ずつ存在していた．それらの F 値は誤認識の多さから非常に小さい値になっている．そこで，その音声を外して相関を出してみたところ，オープンテストデータは正の相関が大きくなり，クローズドテストデータでは無相関であったものが正の相関 (0.6 程度) となった．

表 2 をみると，フィラーの出現位置と音声認識率に関しては，全てのデータにおいて Middle と音声認識率の相関は負の値，Isolate と音声認識率の相関は正の値になった．つまり，フィラーが文中に多いほど音声認識率が悪く，単独で存在しているフィラーの割合が多いほど音声認識率が良くなる傾向が現れた．これは，文中に含まれているフィラーは，前の単語からの接続確率が大きく影響するために予測し難いのではないのかと考えられる．ただし，言語モデル学習データに含まれる文においてフィラーの出現位置を求めてみたところ，Head が 46 %，Middle が 45 %，Isolate が 9 % であった．文頭，文中に偏りなく学習されており，言語モデルの影響で文中に含まれているフィラーが認識され難いわけではなかった．

表 2 より，フィラーの出現位置とフィラー検出率の F 値に関しては，オープンテストデータ以外は Isolate と F 値に正の相関がある．従って，発話に単独で存在しているフィラーの割合が多いほどフィラー検出が良くできる傾向があると考えられる．そこで，実験に用いた全ての音声からフィラーの出現位置別の再現率を調査してみたところ，再現率の高さは順に，単独 (90 %)，文頭 (86 %)，文中 (79 %) となっていた．つまり，単独で存在しているフィラーは，文頭や文中に存在しているものよりも正しく検出されやすい傾向にあると言える．

4. おわりに

本稿では，音声認識を利用したフィラー検出性能と音声認識率との関連性を調査した．

その結果，フィラー検出精度と音声認識率の間には高い相関があることが分かった．また，フィラーの出現位置がフィラー検出性能と音声認識率に影響を及ぼしていることも確認できた．特に，フィラーのみの発話が多いほど音声認識率が良くなり，フィラーを正しく検出しやすい傾向が現れた．従って，フィラーの直前と直後で発話を分割することによって，フィラー検出精度と音声認識率が向上する可能性があると考えられる．

表 2: フィラー検出率・フィラー出現位置の割合・音声認識率の相関

(a) CSJ(クローズドテストデータ)

	F 値	Head	Middle	Isolate	Corr.	Acc.
F 値	1					
Head	-0.72	1				
Middle	0.61	-0.98	1			
Isolate	0.58	-0.26	0.04	1		
Corr.	0.03	0.42	-0.53	0.40	1	
Acc.	0.20	0.29	-0.41	0.44	0.98	1

(b) CSJ(オープンテストデータ)

	F 値	Head	Middle	Isolate	Corr.	Acc.
F 値	1					
Head	-0.05	1				
Middle	0.14	-0.92	1			
Isolate	-0.23	-0.01	-0.40	1		
Corr.	0.37	-0.08	-0.13	0.51	1	
Acc.	0.62	0.05	-0.20	0.37	0.94	1

(c) 講義音声 (28 講義)

	F 値	Head	Middle	Isolate	Corr.	Acc.
F 値	1					
Head	0.53	1				
Middle	-0.69	-0.90	1			
Isolate	0.43	-0.07	-0.38	1		
Corr.	0.82	0.64	-0.76	0.36	1	
Acc.	0.81	0.61	-0.69	0.28	0.99	1

(d) 講義音声 (7 講義)

	F 値	Head	Middle	Isolate	Corr.	Acc.
F 値	1					
Head	0.33	1				
Middle	-0.55	-0.94	1			
Isolate	0.57	-0.24	-0.12	1		
Corr.	0.87	0.29	-0.42	0.35	1	
Acc.	0.82	0.15	-0.28	0.34	0.99	1

参考文献

[1] M.Somiya, K.Kobayashi, H.Nishizaki, and Y.Sekiguchi, “The Effect of Filled Pauses in a Lecture Speech on Impressive Evaluation of Listeners”, InterSpeech 2007, pp.2673-2676, 2007.

[2] 宗宮充宏，西崎博光，関口芳廣，“講義・講演音声を対象としたフィラーの使い方自動評価システム”，音講論 (秋)，2008.

[3] 後藤 真孝，伊藤 克亘，速水 悟，“自然発話中の有声休止箇所のリアルタイム検出システム”，信学論，Vol.J83-D-II，No.11，pp.2330-2340，2000.

[4] 稲垣 貴彦，廣瀬 啓吉，峯松 信明，“話し言葉音声認識における韻律的特徴を利用したフィラー検出”，音講論 (春)，pp.133-134，2008.

[5] 大津展之，“班別および最小 2 乗法に基づく自動しきい値選定法”，信学論，Vol.J63-D，No.4，pp.349-356，1980.