社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＮＬ－168 (21)
2005／7／23

擬似単語モデルによる非言語音声の認識

高橋伸弥† 森元逞†

† 福岡大学工学部 〒814-0180 福岡県福岡市城南区七隈 8-19-1
E-mail: †{takahasi,morimoto}@tl.fukuoka-u.ac.jp

あらまし 従来の音声認識では、咳やくしゃみ、あくびのような非言語音または非音声音は、誤認識を引き起こす雑音として扱われて来た。しかし、自然な音声対話を実現する場合、このような音情報も積極的に利用すべきである。このような非言語音声を検出する方法としては、対象音の音響信号的な特徴を用いた信号処理的アプローチや、対象音から学習した HMM を用いる音声認識的アプローチなどがあるが、対象音の多様性に対処するためには、いずれも多量のデータが必要となる。そこで本研究では、対象となる非言語音声を音素系列で近似表現した疑似単語モデルを提案する。このモデルは、音素認識の結果得られる音素系列をクラスタリングし、上位クラスターの中心となる音素列パターンを非言語音声の近似的な発音とするというものである。提案手法の有効性を確認するために、咳及び咳払いを対象として、音声認識実験を行い、咳／咳払いの波形を学習データとした HMM を用いる手法と比較して、認識正解率、認識精度が改善されることを示した。

キーワード 非言語音声, 擬音語, 咳, 疑似単語モデル

Recognition of Non-Verbal Speech Using Imitated Word Model

Shin-ya TAKAHASHI† and Tsuyoshi MORIMOTO†

† Dept. of EECS, Fukuoka University, Nanakuma 8-19-1, Jonan-ku, Fukuoka, 814-0180 Japan
E-mail: †{takahasi,morimoto}@tl.fukuoka-u.ac.jp

Abstract This paper proposes imitated word models that represent non-verbal sounds, especially cough sounds here, as phoneme sequences. In conventional speech recognition systems, non-verbal sounds, so-called human noises, are processed as burden noises that cause mis-recognition. Non-verbal sounds are, however, important information to know user's physical and psychological condition. In particular, coughing is one of the most important barometers of daily health check, so we propose an approach to detect the cough sounds from user utterances using the imitated word models constructed by clustering of phoneme sequences obtained in phoneme recognition. The experimental results show that this approach can improve the correct rates and the accuracies for words and coughs compared with the approach using HMM constructed from cough waveforms.

Key words non-verbal sound, onomatopoeia,cough recognition, imitated word model

1. はじめに

いわゆる「おと」は音声と非音声に分類することができ、さらに音声は言語音／非言語音、非音声は反射音／物理音に分類することができる [1]。この分類によれば、意識的に発せられるような咳払い、舌打ち、作り笑いなどは非言語音であり、無意識的または反射的に発せられた咳やくしゃみ、笑い声などは非音声 (反射) 音である。従来の音声認識システムでは、このような非言語音／非音声音は誤認識を引き起こす邪魔な雑音として扱われてきた。人間と機械との間での自然な音声対話システムを実現するためには、人間が行っているように、相手の発話内容を理解するだけでなくそれに付随する様々な音情報を総合的に利用することが望ましい。そのためには、言語としての音声情報だけでなく、笑い声や泣き声等の感情を表す音や、咳やくしゃみなどのような身体的な状態を表す音を検出する必要がある。

そこで本研究では、これらの非言語音／非音声音 (以下、非言語音声とする) のうち咳に着目し、ユーザ発話の中から咳を検出する方法を検討する。咳は日常的な健康チェックのための最も重要なバロメータの一つであることから、我々はこれまでに、在宅健康管理音声対話システム [2] におけるユーザの発話中から、咳波形から学習した HMM (以下、非音素 HMM) を用いて、咳を検出することを試みてきた [3]。しかし、このような HMM をベースにした手法では、その性能を高めるために、高

品質な咳データを大量に集めねばならない。特に、咳などのような非言語音声では、自然なデータを大量に収集するのは困難であると予想される。そこで本研究では、このような非言語音声をより簡単に扱うための方法として、擬音語表現に着目し、対象となる非言語音声を音素系列で近似的に表現する方法を検討した。日本語は特に擬音語／擬態語表現が豊富な言語であると言われている [4] が、この理由として、角田は、大脳における優位性レベルの実験結果から、日本人が自然の音や動物の鳴き声など言語音声以外の音も言語機能を司る左脳で処理をしているためであると報告している [5]。本来、言語でない音、音声でない音を、音素言語体系へとマッピングして表現したものが擬音語であるが、本研究は、このような非言語音声を言語的に処理する擬音語的なアプローチを試みたものである。具体的には、咳波形に対する音素認識の結果として得られる音素系列をクラスタリングし、各クラスターの中心となる音素パターンを選択することにより、咳の疑似単語モデルを構築することを考える。この疑似単語モデルは、一般的に使用されている擬音語表現とは異なる表現となりうるが、それをそのまま使用する点が本研究の特徴である。

擬音語認識に関する過去の研究例として、比屋根らは、対象環境音の中心周波数や残響時間から、物を叩く音のような単発音を認識することを試みているが、そのアプローチは信号処理的なものであり言語表現としての擬音語表現を利用してはいない [6]。また他の研究例として、石原らによる環境音の擬音語自動変換の研究があげられるが、彼らのターゲットは主に環境音であり、咳のような非言語音声は扱っていない [7]。また、環境音を音素系列で表現するという点で本研究と似ているが、本研究において、咳の音の多様性を吸収するために複数の疑似単語モデルを用いているのに対し、複数の近似音素をグループ化した音素グループという枠組みにより多様性を吸収している。

本稿では、疑似単語モデルを用いて種々の非言語音声を識別するための第一段階として、咳と咳払いとを識別することを検討する。咳払いは、喉に不快感を感じた場合や、緊張した場合、他人の注意を引きたい場合などに意識的に行う咳だと考えられるが、医学的には咳と咳払いの区別は特に無く、咳払いは軽度の咳症状の 1 種であると言える。そこで、本稿では咳払いを「口を閉じて行う咳」と定義する。

以下では、まず始めに咳および咳払いの音響的な特徴を示した後、疑似単語モデルを用いた非言語音声の認識手法について説明する。次に、咳及び咳払いの波形から構築した HMM を用いた場合との音声認識の比較実験を行い、提案手法が有効であることを示す。

2. 咳及び咳払いの音響分析

図 1 に咳及び咳払いの原波形とそれらに対する音響分析の結果を示す。これらの波形は サンプリングレート 16kHz、量子化ビットレート 16bit で比較的静かな室内で録音されたものである。以下で使用する全てのデータセットは同一条件で録音された。上段の波形は原波形を示し、中段はパワー包絡、下段はサウンドスペクトログラムである。

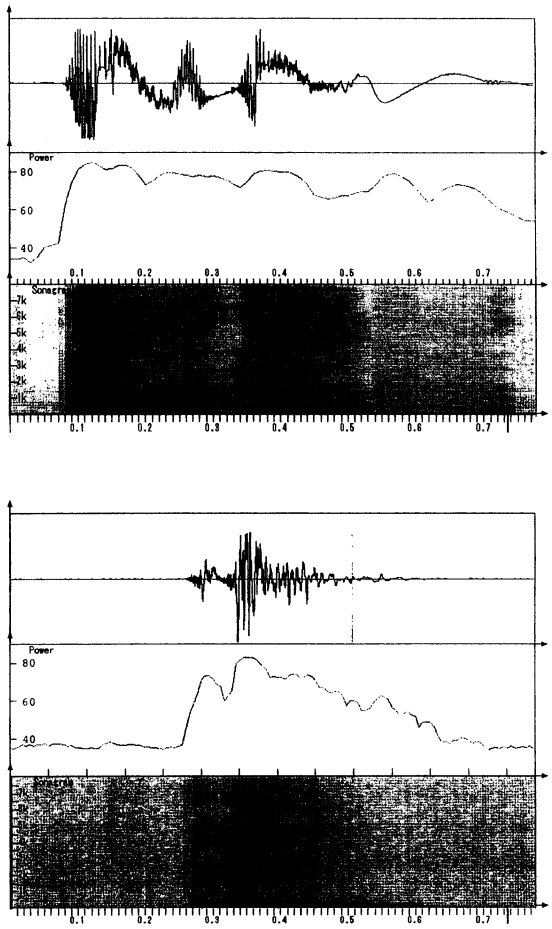

図 1　咳波形 (上) 及び咳払い波形 (下)

図 1 に示すように、咳波形においては急激な立ち上がりのあと一気に空気を流出するため、咳払いと比べて激しい音圧変化が見られる。またスペクトログラムより、ピークの見られる周波数帯にも違いがあることがわかる。

3. 疑似単語モデルの作成方法

3.1 基本的な考え方

非音素 HMM を用いて高精度に咳を検出するためには、高品質かつ大量の咳データを集めるだけでなく、適切なモデルを構築するために咳の音響的な特徴や HMM のトポロジカルな構造を検討する必要があり、より多様な咳を対象とした場合、さらに難しくなるだろうと思われる。そこで、我々は咳を HMM で表現する代わりに音素系列として表現する疑似単語モデルを提案する。基本的に、この考え方は、咳を擬音語としてモデル化することと同等である。しかし、咳には様々なヴァリエーションがあり、石原らが同様の問題を指摘しているように、擬音語の表現は一意ではない [7]。例えば、咳の擬音語としては、「ゴホゴホ」や「ゲホッゲホッ」「コホンコホン」等、様々なものが考えられる。上記のような一般に用いられている擬音語表現

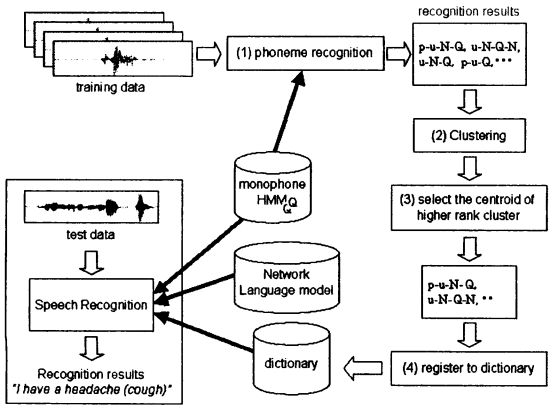

図 2　モデル化の処理の流れ

は、文化や慣習のなかで既に記号化されたものであるから、これらの表現をそのまま用いて咳を認識することは難しいと考えられる。

そこで本研究では、音素認識とクラスタリングを用いて、咳を音素系列として近似することを考える。

具体的には、

（1） 音節構造を持たない任意の日本語音素列を受理するような文法と日本語モノフォン HMM を用いて学習データを認識し

（2） 認識結果の音素列をクラスタリングして上位クラスタの中心パターンを音素列パターンとして選択し

（3） これらを咳の音素列パターンの発音として辞書に登録する

という処理を行う。図 2 にこの処理の流れを示す。咳及び咳払いに対して、この処理を実行することにより、それぞれに対する疑似単語モデルを得ることが出来る。

3.2 音素認識による非言語音声の認識

まず始めに、日本語モノフォン HMM を用いて音素認識を行った。表 1 は、この実験に用いた日本語音素である。表 1 の "N" は撥音であり、"Q" は促音を表す。ここで用いた音素認識では、日本語の音節構造、すなわち、子音の後に必ず母音が続くといった規則や、撥音及び促音が語頭に来ることはない、といった日本語の規則は用いていない。

学習データは、4 人の男性による 120 の咳と、3 人の男性による 118 の咳払いである。モノフォン HMM には、大語彙日本語連続音声認識ツール Julius [8] に付属のものを使用した。

表 1　実験で用いた日本語音素

vowels	/a/,/e/,/i/,/o/,/u/
long vowels	/a:/, /e:/, /i:/, /o:/, /u:/
consonants	/b/, /by/, /ch/, /d/, /dy/, /f/, /g/, /gy/, /h/, /hy/, /j/, /k/, /ky/, /m/, /my/, /n/, /ny/, /p/, /py/, /r/, /ry/, /s/, /sh/, /t/, /ts/, /w/, /y/, /z/
moraic silence	/Q/
moraic nasal	/N/

表 2　音素認識の結果の例

咳	/u/-/Q/
	/u/-/f/-/u/
	/b/-/u/-/Q/
	/u/-/f/-/Q/-/N/
	/z/-/g/-/f/-/Q/
	/p/-/i/-/Q/-/f/-/u/
咳払い	/p/-/u/-/N/
	/u/-/N/-/Q/
	/u/-/N/-/i/-/Q/
	/z/-/h/-/Q/-/N/
	/p/-/u/-/N/-/Q/
	/p/-/u/-/N/-/z/-/Q/

表 3　高頻度に現れた音素

	/u/	/Q/	/f/	/z/	/p/	/a/	/i/
咳	108	43	27	15	15	11	10
	/u/	/N/	/Q/	/p/	/i/	/z/	/e/
咳払い	93	62	56	51	29	27	25

表 4　辞書に登録した疑似単語モデル

咳	/f/-/u/-/Q/
	/u/-/Q/
	/z/-/u/-/Q/
	/u/-/f/-/u/-/Q/
咳払い	/p/-/u/-/N/-/Q/
	/u/-/N/-/Q/-/N/
	/u/-/N/-/Q/
	/p/-/u/-/Q/

音素認識の結果を表 2 に示す。さらに表 3 に、頻度の高い順に 1 位から 7 位までの音素を示す。咳及び咳払いに対する平均音素数は、それぞれ 2.3 及び 3.4 であった。咳払いに対する平均音素数が咳よりも大きいのは、咳払いの方が、持続時間が短いと思われる子音の頻度が高いからであると考えられる。表から、咳及び咳払いの両方で、母音 /u/ 及び 促音 /Q/、無声破裂音 /p/ が高頻度で現れていることが分かる。また、摩擦音 /f/, /z/ が咳に対しては、高頻度で現れているのに対して、咳払いに対しては、撥音 /N/ が高頻度で現れているのが特徴的である。

3.3 クラスタリングによる疑似単語モデルの決定

次に、音素認識で得られた音素系列をクラスタリングによって分類し、クラスタの大きさが 1 位から 4 位までのものを選び、その中心パターンを対象となる非言語音声の近似的な発音として辞書に登録する。クラスタリングの手法には、最大距離法 [9] を用いることとし、パターン間距離は以下の式で計算した。

$$d(A,B) = 1 - \frac{\mathrm{num}(A \cap B)}{\mathrm{num}(A \cup B)} \qquad (1)$$

ここで、A 及び B はそれぞれのパターンに含まれる音素記号の集合であり、$\mathrm{num}(X)$ は集合 X に含まれる要素数である。

距離の定義は上記以外にも考え得るが、音素パターンの長さ

表 5　受理される文パターンの例

(咳)-(咳)-頭-も-(咳)-ちょっと-痛い-ん-です-けど-(咳)
体調-は-(咳)-(咳)-少し-悪い-です-(咳)-(咳)
咳-が-かなり-出-ます-(咳)-(咳)-(咳)-(咳)

表 6　認識結果

	全体		単語	
	%Corr.	%Accu.	%Corr.	%Accu.
疑似単語モデルを用いた手法	86.2	74.1	93.3	91.1
非音素 HMM を用いた手法	66.4	29.2	95.3	93.2
疑似単語モデル＋後処理	92.8	78.9	93.5	91.2
非音素 HMM ＋後処理	79.2	56.5	95.3	92.7
	咳		咳払い	
	%Corr.	%Accu.	%Corr.	%Accu.
疑似単語モデルを用いた手法	73.6	60.8	85.3	8.3
非音素 HMM を用いた手法	23.6	8.2	30.3	-341.3
疑似単語モデル＋後処理	94.0	73.4	85.7	0.95
非音素 HMM ＋後処理	40.8	5.4	31.4	-112.4

が比較的短いことから、ここでは集合間の距離を用いることとした。

3.4　疑似単語モデルの辞書への登録

以上の方法から求めた疑似単語モデルを表 4 に示す。それぞれのパターンは非常に類似しているが、咳は摩擦音 /f/,/z/ を用いて表現されているのに対し、咳払いでは撥音/N/が用いられていることが分かる。

4.　音声認識実験

4.1　咳及び咳払いを含んだ発話の認識

咳及び咳払いに対する非音素 HMM を用いた手法と疑似単語モデルを用いた手法とを比較するために音声認識実験を行った。比較のために用いられた 咳及び咳払いの非音素 HMM はそれぞれ、3 状態の Left-to-Right 型 HMM であり、疑似単語モデルを作成するのに用いたものと同じデータを用いて学習した。特徴パラメータには、12 次元 MFCC ＋ 12 次元 Δ MFCC ＋ Δ 対数パワーの計 25 次元を用いた。HMM の学習及び音響分析、音声認識には、HTK [10] 付属のプログラムを使用した。実験に用いたテストデータは、咳払いを含む音声発話 58 (男性 2 名)、咳を含む音声発話 120 (男性 4 名) であり、のべ単語数 751、咳回数 416, 咳払い回数 109 となっている。また言語モデルには、[2] で使用している症状の訴えに関する文を受理するネットワーク言語モデルに、文頭、文末及び文節間での咳または咳払いを追加したものを用いた。ここで、咳または咳払いは任意の回数の継続を許している。この言語モデルは、受理される文パターン数が約 50、語彙数は約 200 であり、具体的には、表 5 のような文を対象としている。

4.2　認識結果と考察

表 6 に実験結果を示す。この表では、咳の挿入誤りの影響をみるために、単語及び咳の正解率及び認識精度を別々に計算している。さらに、咳の回数を知ることよりも咳の有無を知ることが重要であると考えて、認識結果内の連続した複数回の咳を 1 回の咳として置き換える後処理を行った。この後処理を施した結果を表 6 に併せて示す。

表から分かるように、提案手法により、非音素 HMM を用いる方法と比較して、単語に対する正解率と認識精度は若干低下しているものの咳及び咳払いに対する正解率が大幅に改善されている。認識精度に関しては、提案手法においても非常に低い値となっているが、それでも非音素 HMM の場合と比べると改善することが出来た。後処理によってさらに正解率が改善できているが、認識精度がやや低くなっている。後処理における認識精度の低下の理由は、挿入誤りの数だけでなく認識精度の分母である咳及び単語の正解数が減少したためである。

表 7　誤認識の詳細

	疑似単語モデル	非音素 HMM
正解認識数		
単語	701	716
咳	306	98
咳払い	93	33
置換誤り		
咳 → 単語	1	4
咳払い → 単語	0	0
単語 → 咳	3	1
単語 → 咳払い	14	3
咳 → 咳払い	27	301
咳払い → 咳	15	74
単語 → 単語	28	29
削除誤り		
単語	5	2
咳	82	13
咳払い	1	2
挿入誤り		
単語	17	16
咳	53	64
咳払い	83	405

表 8　単語との間での置換誤り

疑似単語モデル	頻度
/u/-/N/ → 咳払い	12
/s/-/u/-/k/-/o/-/sh/-/i/ → 咳	2
/g/-/a/ → 咳払い	1
/m/-/u/-/n/-/e/ → 咳払い	1
/s/-/e/-/k/-/i/ → 咳	1
咳 → /N/	1
非音素 HMM	頻度
咳 → /u/-/N/	2
咳 → /N/	2
/g/-/a/ → 咳払い	1
/m/-/u/-/n/-/e/ → 咳払い	1
/s/-/e/-/k/-/i/ → 咳払い	1
/u/-/N/ → 咳払い	1

4.3　誤認識の解析

表 7 に音声認識実験の誤認識の詳細を示す。この表から分かるように、非音素 HMM を用いた方法では、咳払いの挿入誤りの数と、咳を咳払いと誤認識する置換誤りの数が、非常に大き

くなっている。疑似単語モデルを用いた方法でも、咳払いの認識精度がかなり低いものとなっているが、これは非音素 HMM の場合と同様、データの録音時における環境雑音により、無音区間を咳払いと誤認識したためだと考えられる。これを解決する方法としては、咳／咳払いの挿入を抑制するようなペナルティの付与などが考えられる。

一方、2種の咳の間での置換誤りについて見てみると、提案手法では、非音素 HMM を用いる方法と比べて、1/10 以下に低減できていることが分かる。これらの結果から、提案手法により咳の識別性能が大幅に改善できることを確認できた。

単語とそれぞれの咳との置換誤りの詳細を表 8 に示す。この表から、提案手法により、"うん"(/u/-/N/) を咳払いと誤認識する例が最も多かったことがわかる。"うん"(/u/-/N/) と正しく認識した個数は 17 であることから、全体の 41% が咳払いとして誤認識したことになる。これは、咳払いの発音として疑似単語モデル "/u/-/N/-/Q/" を辞書に登録したためであるが、疑似単語モデルを用いる際には、既存の読みと重複しないよう注意しなければならない。

5. おわりに

本稿では、疑似単語モデルを用いてユーザ発話から咳を検出する方法を提案した。非音素 HMM を用いる手法と比べて、より高い正解率・認識精度を得ることが出来た。また2種の咳の識別性能も改善した。非音素 HMM を用いた手法が十分な性能に達していないのは、学習データ量が十分では無かったためだと考えれるが、逆に言えば、提案手法は、同様の条件下、すなわち学習データ量が十分では無い場合でも、比較的良い性能を発揮できると言える。十分な量の学習用データの収集が難しいと思われる咳のような非言語音声に対しては、本手法が有効であると思われる。

今後の課題として、提案手法を他の種類の咳、例えば一般に診断に用いられている乾性咳及び湿性咳に適用すること、また咳症状の軽重を判別することが考えられる。さらに、疑似単語モデルを用いた他の非言語音声の検出も検討したい。

謝辞 本研究は、2004 年度文部科学省の科学研究費補助金 (若手研究 (B) No. 16700195) の支援により行われた。

文　献

[1] 風間喜代三, 松村一登, 上野 善道, 町田 健, "言語学", 東京大学出版会, 1993.

[2] S. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta, "Dialogue Experiment for Elderly People in Home Health Care System", in *Lecture notes in computer science*, Springer (TSD 2003 Proc), 2807, pp.418–423, 2003.

[3] S. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta, "Cough Detection in Spoken Dialogue System for Home Health Care", Proc. of the 8th International Conference on Spoken Language Processing, pp.1865–1868, Jeju Island, Korea, Oct. 2004.

[4] 田守育啓, ローレンス・スコウラップ, "オノマトペ – 形態と意味 –", くろしお出版, 1999.

[5] 角田忠信, " 日本人の脳 – 脳の働きと東西の文化 –", 大修館書店, 1978.

[6] 比屋根一雄, 澤部直太, 飯尾淳, "単発音のスペクトル構造とその擬音語表現に関する検討", 電子情報通信学会音声研究会, 信学技法, SP97-125, pp.65-72, 1998.

[7] 石原一志, 駒谷和範, 尾形哲也, 奥乃博, "環境音を対象とした擬音語自動認識", 人工知能学会学会誌 Vol.20, No.3, pp.229–236, Mar. 2004.

[8] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine", Proc. of the 7th European Conference on Speech Communication and Technology (EuroSpeech), pp. 1691–1694, Aalborg, Denmark, Sept. 2001.

[9] 長尾真, "パターン情報処理", 電子情報通信学会編, コロナ社, 1983.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book", Cambridge, U.K., HTK version 3.0 edition, 2000.