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

2008－MUS－74 (22)
2008－SLP－70 (22)
2008／2／9

発話の困難な障害者のための声質変換・HMM音声合成を用いた日英音声合成システムの構築

加島 慎平[*1] 飯田 朱美[*2] 安 啓一[*1] 相川 恭寛[*3] 荒井 隆行[*1] 菅原 勉[*1]

*1 上智大学 *2 東京工科大学

*3 株式会社ループドピクチャー

概要

本報告では発話の困難な日本人 ALS 患者のコミュニケーション支援を目的として、日英両言語バイリンガル音声合成システムの構築を行った。本研究では HMM 音声合成ツールキット HTS を用いて、患者の声質を反映した HMM に基づく日本語音声合成システムを構築した。また、声質変換を用いた方法も提案し、音声合成を実現した。本報告では、構築した日本語音声合成システムについて、手法や構築にかかる患者の録音量などの面からシステムの評価を行った。また、同様に患者の声質を持つ英語音声合成用の HMM 音響モデルも構築した。そして日英両言語の音声合成システムを GUI によって統合し、患者のための日英バイリンガル音声合成システムを完成させた。

Development of a Japanese and English Speech Synthesis System Based on HMM Using Voice Conversion for the People with Speech Communication Disorder

Shimpei Kajima[*1], Akemi Iida[*2] Keiichi Yasu[*1] Yasuhiro Aikawa[*3]
Takayuki Arai[*1] and Tsutomu Sugawara[*1]

*1 Sophia University *2 Tokyo University of Technology

*3 Looped Picture Company

Abstract - In this paper, we describe our work on developing a bilingual communication aid for a Japanese amyotrophic lateral sclerosis (ALS) patient. For this study, HTS toolkit was used to develop the system. First developed was a Japanese speech synthesis and two methods were attempted. The first used an acoustic model built from the recordings of the patient's read speech. The second used an acoustic model built from a voice converted to the patient's voice. The result of the perceptual experiment showed that the voice synthesized with the latter was perceived to have a closer voice quality to the patient's natural speech. An acoustic model for English speech synthesis system was built and GUI works on windows was developed for the patient to use.

1. はじめに

コミュニケーションの手段として音声は非常に重要である。しかし、呼吸器官など発話に必要な器官の障害のために音声によるコミュニケーションが容易ではなかったり、不可能であったりする障害者も多く存在する[1]。発話が困難となる病気の一つとして筋萎縮性側索硬化症（Amyotrophic lateral sclerosis: ALS）[1]が挙げられる。ALSは、思考、知覚、感覚能力は正常でありながら、徐々に筋肉が動かなくなり、やがては全身麻痺となる難病である。麻痺の進行は人それぞれであるが、呼吸筋の麻痺にまで症状が進行すると、人工呼吸器の装着のため、気管切開を余儀なくされるケースが多い。現在、人工呼吸器の装着による在宅人工呼吸療養が可能になったが、気管切開後の発話は大変困難であり、このような意思伝達の術が著しく限定された状況下で、「生活の質 (Quality of life: QOL)」をいかに向上できるかは大きな課題である[2]。患者からも、コミュニケーションがとれないことが一番つらいと言う声が多く寄せられている[1]。

ここで本研究の協力者である故山口進一氏を紹介する。

本稿では山口氏の生前からの要望により実名と個人情報を記す。山口氏は1996年、58歳の時にALSを発症、2006年5月に67歳で逝去された。山口氏はALS発症後もパソコンの利便性についての講演活動を精力的に行っていた。そして、声を失った後も自分の声で話したいという強い要望を持っており、1999年以降、飯田とともに協力者として研究に参加して下さった。本研究では以降山口氏を協力者と記す。

音声によるコミュニケーション支援システムとして、従来から音声合成が利用されてきた。特に、コンピュータに打ち込んだテキストを規則合成方式[3]によって合成し出力するTTS (Text-to-Speech)音声合成システムは発声器官の障害等により発話能力の低下した患者のコミュニケーション補助システムとして非常に有効である[4]。しかし、従来のシステムでは合成音声が機械的で、かつ自分本来の声ではないという問題点があった。協力者は自らの声を使ったコミュニケーションを強く希望していた[5]。そこで飯田らは、録音した音声をデータベースとするコーパスベース型音声合成システムを用いることで、協力者の録音音声を元に協力者の声質を持った日本語コーパスベース型音声合成システム構築を行った[4]。その後、講演活動での必要性から、協力者から英語の合成音声システムも構築できないかという要望があった。そこで、協力者の録音音声から英語音声合成システムを構築した。コーパスベース型音声合成システムを構築する上で大きな問題となったのは、合成用データベースを構築するために必要な協力者の録音音声の量が膨大であるという点である。我々は声質変換技術[6]を用いて、2000年に収録した協力者のTIMIT[7]の発話246文から協力者の声質を学習することで、協力者の声質をもつ音声データベースを構築し、英語音声合成を実現した[8]。前述のコーパスベース合成方式による日英両言語の音声合成システムを用いて、バイリンガルシステムの試作を行った[9]。

本報告では、上記の研究過程を踏まえた次の段階として、隠れマルコフモデル（HMM）に基づく音声合成[10]を導入した日英バイリンガル音声合成システムについて述べる。初めに、オープンソースプログラムのHMM音声合成ソフトウェアを用いて、協力者のHMMに基づく日本語音声合成システムの構築を試みた。また、声質変換を用いることで、より少ない録音量でのシステムの構築方法を提案した。同様に声質変換を用いて患者の声質を反映した英語音声合成システムも構築し、日英バイリンガル音声合成システムを実現した。本研究で提案する手法で音声合成システムを構築することで、発話の困難な患者に大きな負担を強いることなく手軽に患者自身の声質を持つ音声合成を提供することが可能になり、コミュニケーションの円滑化が期待できる。

2. 協力者の声質を持つ日本語HMM音声合成システムの構築

2.1 HMM音声合成の導入

先行研究[5,10]で試作したシステムでは、コーパスベース音声合成方式を用いた。この方法では本人の声質をそのまま合成音声に反映できる利点があるが、すべて本人の音声を録音したものから音声合成用のデータベースを構築しなければならず、患者の発声にかかる負担が大きいことが問題であった。そこで、我々は次のステップとしてオープンソースのHMM（hidden Markov model）音声合成ソフトウェアを用いて日本語音声合成システムの構築に取り掛かった。HMM音声合成用の音響モデルの構築にはHTS（HMM-based Speech Synthesis System）[11]を用いた。これは、HMM音声認識用ツールキットHTK (hidden Markov model toolkit)[12]を音声合成システムツールとして拡張するパッチプログラムであり、一定量の文章の読み上げ音声とそのラベルファイルから発話者に固有のHMMモデルを学習し、音声合成用の音響モデルを作成する。この手法の導入により、先行研究で用いたダイフォン合成方式の音声合成ではなく、より自然な音声の合成が期待できる。

2.2 HMMに基づく日本語音声合成システム

HTSの出力する音響モデルに対応する日本語音声合成ソフトウェアGalatea Talk[13]を用いて、協力者の声質を持つ日本語音声合成システムの構築を試みた。HTSによる音響モデルの学習量としては現在のところATR503文[14]の全文を録音することが妥当であるとされている。しかし、我々が目指す音声合成システムの使用者は発話の困難な障害者であり、そういった使用者にとって、503文章の読み上げは大きな負担となる。本報告の協力者についても2000年にATR Bセットの収録を先行研究[4]のために行ったが、本人の負担を考え、すべてのバイフォンが出現するように全503文の中から129文を選定し、最終的に108文を録音することができた。このように、音声合成システムの構築に503文の収録は対象の使用者の負担が大きく、この収録量では断念せざるを得ない対象者も少なくないはずである。また、129文に減らしたとしても依然として負担は大きい。そのような背景から、まず、協力者の発話した25文を用いてHTSによる音響モデル（yam25）の構築を試みた。25文の根拠は、音声認識・音声対話技術講習会（京都大学学術情報メディアセンター主催、2007）で最少収録量が20文だったた

め、それよりも少し多い量とした。そして、協力者の発話すべて（108 文）を用いて学習した音響モデル（yam108）も作成した。その結果、声質としては協力者の話者性を感じるものの、yam25 で合成した音声は自然性を大きく欠き、yam108 による合成音声も自然性において満足の行く結果には至らなかった。原因の一つとして、今回の実験では、ラベルデータの修正を自分たちで行うには至らず、標準のラベルデータを用いたことが考えられる。よって、ラベルデータの修正ができれば、高品質な音声を合成できると考えられる。今後、ラベルデータの整備を含めて再検討していきたい。

2.3 声質変換の利用

先行研究[8][15]により、声質変換技術を用いることで、患者の話者性を反映したままより少ない録音量で音声合成システムを構築できることがわかった。そこで上記の HTS を用いた日本語音声合成システムにも声質変換の利用を試みた。声質変換利用前のシステムと利用後のシステムの概略図を図 1 に示す。HMM によって構築される音響モデルの話者性、自然性は学習文章量に大きく影響を受ける。そこで、以下の手順で協力者の声質をもつ ATR B セット 503 文全文を得た。

1. HTS にあらかじめ用意された男性話者の発話（ATR503-m001）による 503 文全文とそのラベルデータを用意
2. 男性話者と協力者の 503 文から共通な発話 25 文を選定し、声質を学習
3. 声質変換によって男性話者の 503 文を協力者の声質に変換

変換後の 503 文で HTS を用いて音響モデルを学習させることで、協力者の発話 25 文から 503 文で学習した音響モデルを得ることができる。この方法により協力者の話し方などの F0 以外の韻律情報はモデル化されないが、高い自然性を持った音声合成が期待できる。また、ラベルデータは標準のものを利用できるという利点もある。構築した音響モデルの詳細を表 1 にまとめる。

2.4 日本語音声合成システムの評価

構築した録音音声の HTS および声質変換後音声の HTS の 2 種類の方法による協力者の日本語合成音声について、協力者の録音データとの比較評価実験を行った。

刺激に用いた音声は以下の 5 種類とした。

- yamorg: 協力者の録音音声
- yam25: ATR B セット 503 文から協力者の発話した 25 文を用いて HTS で音響モデルを学習した合成音声
- yam108: 503 文から協力者の発話した 108 文を用いて HTS で音響モデルを学習した合成音声
- yam25vc503: 503 文から協力者の発話した 25 文を用いて声質を学習し、声質変換を行った 503 文で HTS による音響モデルを学習した合成音声
- yam25vc108: 503 文から協力者の発話した 25 文を用いて声質を学習し、声質変換を行った 108 文で HTS による音響モデルを学習した合成音声

刺激文には協力者が日常的に使用する頻度が多い文章から 6 文選定し、GalateaTalk で合成した。使用した文章を以下に示す。

- エアコンをつけて下さい。
- 体の向きを変えてください。
- おはようございます。
- 元気にしています。
- お世話になりました。
- 御苦労さまでした。それでは失礼します。

すべての音声は 16 kHz でサンプリング、16 bit で量子化された。音響モデル構築の処理条件を表 2 に示す。

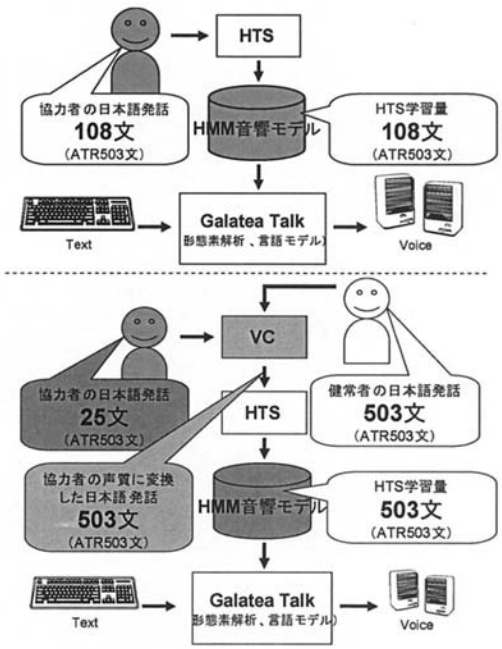

図 1. 日本語音声合成システム
(上：HTS で構築、下：声質変換後、HTS で構築)

表 1. 構築した音響モデルとその詳細

音響モデル名	Voice Conversion			HTS	
	ソース	ターゲット	学習文章量	話者	学習文章量
yam25	-	-	-	yam(自声)	25 文
yam108	-	-	-	yam(自声)	108 文
yam25vc503	ATR-m001	yam(自声)	25 文	yam(変換)	503 文
yam25vc108	ATR-m001	yam(自声)	25 文	yam(変換)	108 文
yam108vc503	ATR-m001	yam(自声)	108 文	yam(変換)	503 文

表 2. 音響モデル構築の処理条件

ソース	ATR-m001
VC 学習文章	ATR503
VC 学習量	25 文・108 文
GMM クラス数	32
VC での MFCC 次数	24
HTS 学習文章	ATR503
HTS 学習量	25～503 文
HTS での MFCC 次数	18
HMM 状態数	5
標本化周波数	16 kHz
分析フレーム長	25 msec
シフト長	5 msec

20 代から 40 代の男女を実験参加者とした。実験は Web 上で行い、実験環境は参加者に任意のヘッドフォン着用を条件とした以外は実験参加者の任意とした。実験参加者には yamorg を聞いてもらい、残りの 4 刺激を聞き、yamorg との音声の類似度を 5 段階で評価（1: まったく似ていない、5: 非常に似ている）してもらった。音声は何度でも聞いてよいものとした。この過程を刺激文ごとに（計 6 セット）行なってもらった。

2.5 結果・考察

類似度の平均値と標準偏差を図 2 に示す。$\chi 2$ 検定の結果、全体として有意に差があるという結果が得られた。声質変換を用いた yam25vc503 と yam25vc108 の 2 刺激が、HTS のみの刺激である yam25 と yam108 の類似度を上回り、声質変換を利用したほうが協力者本人の発話に近いことが確認できた。

システムを使用する患者の負担となる、学習に必要な録音量に注目すると、yam25 と yam25vc503 は共に必要とする患者の録音量が 25 文である。2 つの刺激の類似度を比較すると、yam25vc503 のほうが高い値を示している。また、協力者の長女の高橋雅子さんにも聴取実験にご協力いただき、その結果も実験結果の平均と同様に yam25vc503 のスコアが yam25 よりも高かった。このことから、同じ少量の録音量で音声合成システムを構築するのであれば、声質変換を用いる方が、患者の元の声に近い音声を合成できることがわかった。HTS から音響モデルを構築するのに必要な学習量に注目すると、yam108 と yam25vc108 は HTS での学習量がともに 108 文である。二つの類似度に大きな差は無く、また、HTS の学習量が等しいため、合成音声の自然性の点では大きな差が無いと考えられる。よって、声質変換での学習量が 25 文でも、録音音声で構築した音響モデルと同様の声質を持った音声が合成できることがわかった。しかし、今回は協力者の 108 文の発話にラベルデータは存在しないため、yam25、yam108 では適当な音響モデルが学習できているとは一概には言えない。しかし、ラベルデータの付与は手作業によるところが大きいため、音声合成システムの構築の負担にもなる。そういった意味では、声質変換を利用した方法も本研究の対象者のような場合には十分選択肢となりえると考えられる。

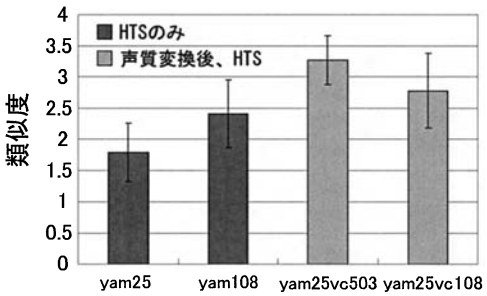

図 2. 各刺激と yamorg との類似度

3. 協力者の声質を持つ英語音声合成システムの構築

3.1 英語話者のダイフォンデータを声質変換して構築した英語音声合成システム

先行研究[8]では Festival[16]に内蔵された英語話者 kal の合成音声との声質変換により、協力者の声質を持つ英語音声合成用のダイフォンデータベースを得た。その結果、合成音声に協力者の話者性が反映されたことが確認できたが、多くの研究者から合成音声の音質が協力者の録音音声に比べて「こもった」と感じたという意見を得

た。そこで、前章で構築した日本語音声合成と同じように、英語話者の録音音声と協力者の録音音声との間で声質変換の学習を行った。Carnegie Mellon 大学の John Kominek 氏の協力を得て、Kominek 氏の Festival のダイフォンデータベース（jmk）と TIMIT 発話の一部（100 文）を頂いた。Kominek 氏と先行研究[4]での音声収録時に録音した協力者の TIMIT 発話から共通の 27 文を選び出し、声質の学習を行った。学習した変換関数を用いて Kominek 氏の 1369 パターンのダイフォンを協力者の声質に変換し、変換したダイフォンを用いて Festival による音声合成を行った。合成した音声を聴取し評価したところ、声質は協力者の声質を良く反映していると感じた。しかし、音質に関してはぶつぶつと細かく音声が切れるような雑音が生じた。声質の学習に用いた Kominek 氏と協力者の英語発話の間でセグメンテーションずれが生じていたこと、また、jmk ダイフォンデータベース自身もまだセグメンテーションの整備が十分ではないことなどが要因と考えられ、今後の改善は期待できる。

3.2 声質変換と HTS を用いた協力者の HMM 音響モデルの構築

前章では、声質変換を利用した HMM に基づく日本語音声合成が、発話の困難な患者のための音声合成システム構築に有効であることを述べた。同様に、声質変換と HTS を用いて協力者の声質を持つ、Festival 用の音響モデルの構築を行った。

HTS が Festival 用音響モデルの構築に必要な学習文章量は CMU ARCTIC[17]の全 1132 文が妥当とされている。そこで、前節で学習した変換関数を用いて、公開されている Kominek 氏の発話した CMU ARCTIC の a セット 593 文を協力者の声質に声質変換することで、協力者の声質を持った CMU ARCTIC を得ることができ、これを用いて Festival 用音響モデル yam27vc593ENG を構築した。yam27vc593ENG 構築の処理条件を表 5 に示す。今後は構築した音響モデルを元に Festival による英語音声合成の実現が目標となる。

4. 日英両言語の HMM 音声合成を用いたバイリンガルシステムの構築

3 章で構築した日英両言語の協力者の声質を反映した音声合成システムを Windows 上で動かすためのグラフィカルユーザインタフェースの構築が行った。インタフェースの概観を図 3 に示す。Festival は Linux 上で動作するため、以前は VMware [18]を用いて Windows 上で仮想 OS として Linux を動作させ、Linux 上で GUI を用いることを検討した。しかし、今回作成した GUI は Cygwin [19]を用いて Festival を Windows 上で動かすことで、仮想 OS を起動しなくても Windows 上でスタンドアロンのシステムを動作させることが可能になった。これにより、コンピュータを使用する患者の多くが用いている Windows 上でバイリンガルシステムを利用することができ、より多くの患者に受け入れられやすいシステムとなったと考えている。

表 5. 英語合成用音響モデル yam27vc593ENG の処理条件

ソース	jmk
VC 学習文章	TIMIT
VC 学習量	27 文
GMM クラス数	32
VC での MFCC 次数	24
HTS 学習文章	CMU ARCTIC a set
HTS 学習量	593
HTS での MFCC 次数	18
HMM 状態数	5
標本化周波数	16 kHz
分析フレーム長	25 msec
シフト長	5 msec

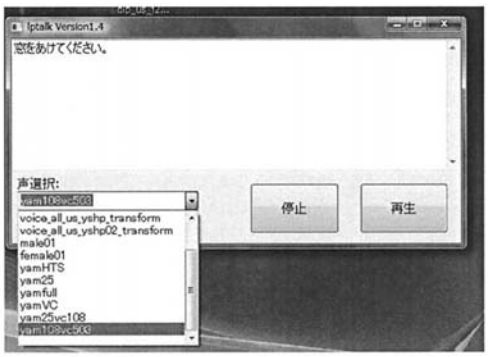

図 3. バイリンガルシステムの GUI

5. まとめ

日本人 ALS 患者の録音音声を元に、HTS を用いて患者の声質を持った HMM に基づく日本語音声合成システムの構築を行った。患者の発話した 108 文章から HMM 音声合成ツールキット HTS を用いて音響モデルを構築した。また、システムの構築に必要な多量の音声録音が、発話の困難な患者の大きな負担となることから、声質変換を利用し、患者の声に声質変換した健常者の十分な量の発話とラベルデータから音響モデルを学習した。患者の発話から学習した音響モデルと声質変換後の発話から学習した音響モデルの 2 種類のモデルについて聴取実験

による評価を行った。その結果、患者の発話量が少ない場合、声質変換を用いたほうが患者の自声に近い音声を合成できることが確認できた。同様に、声質変換を用いたHMMに基づく英語音声合成用音響モデルも構築した。日英両言語のHMM音声合成システムをGUIによって統合し、Windows上でスタンドアロンに動作する協力者の声質を反映したバイリンガル音声合成システムを構築した。

謝辞

本研究は科学研究費補助金（A-2, 16203041）の助成を受けて行った。録音に協力してくださった故・山口進一氏とそのご家族の方々、Festival利用面、音声提供でご協力して下さったCarnegie Mellon 大学のJohn Kominek氏、HTSについてご教示頂いた、名古屋工業大学の酒向慎司先生、全炳河先生に感謝申し上げます。本研究の一部は、文部科学省私立大学学術研究化推進事業上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の支援を受けて行った。

参考文献

[1]豊浦保子, “生命のコミュニケーション･筋萎縮性側索硬化症患者の記録”, 東方出版, 1996.

[2] 日本ALS協会ホームページ：ALSの症状と生活の変化； Retrieved from http://www.alsjapan.org/contents/whatis/02.html

[3]古井貞熙, “音響・音声工学”, 近代科学社, 1992.

[4] A. Iida and N. Campbell, “Speech database design for a concatenative text-to-speech synthesis system for individuals with communicative disorders”, International Journal of Speech Technology 6, pp. 379-392, 2003.

[5] 山口進一, “パソコンを使いこなそう”, 日本 ALS 協会福岡支部第四回総会記念講演, Retrieved from http://www.ne.jp/asahi/laconicmako/ikiru/toukou/pasokon_tukaikonasou.pdf, 1999.

[6] D. Childers, B. Yegnanarayana and Ke Wu, “Voice conversion: factors responsible for quality”, IEEE International Conference on ICASSP '85, Vol. 10, pp. 748-751, 1985.

[7]J. Garofolo, L. Lamel, W. Fisher, et al., “Darpa timit acoustic-phonetic continuous speech corpus”, Technical report NISTIR 4930, National institute of standards and technology, Gaithersburg, MD, 1993.

[8] 加島慎平, 飯田朱美, 安啓一, 荒井隆行, 菅原勉, “声質変換技術を用いた日本語話者のための英語音声合成システムの構築”, 日本音響学会秋季研究発表会講演論文集, pp. 251-252, 2006.

[9] 飯田朱美, 加島慎平 他, “ALS 患者のためのバイリンガル音声合成システムの構築と評価”, 日本音響学会春季研究発表会講演論文集, pp. 341-342, 2007.

[10] A. Falaschi, M. Giustiniani and M. Verola, "A hidden Markov model approach to speech synthesis," Proc. Eurospeech '89, Vol. 2, pp. 187-190, 1989.

[11] K. Tokuda, et al., “The HMM-based speech synthesis system (HTS) “, Retrieved from http://hts.ics.nitech.ac.jp/.

[12] Univ. of Cambridge, HTK homepage, Retrieved from http://htk.eng.cam.ac.uk/.

[13] 嵯峨山茂樹, 川本真一, 他, “擬人化音声対話エージェントツールキット Galatea “, 情報処理学会研究報告, 2002-SLP-45-10, pp. 57-64, 2003.

[14] 阿部匡伸, 匂坂芳典, 他, “研究用日本語音声データベース利用解説書 (連続音声データ編) “, ATR自動翻訳電話研究所, 1990.

[15] 加島慎平, 飯田朱美, 安啓一, 荒井隆行, 菅原勉, “声質変換機能を用いた日本語話者のための英語合成音声の了解度評価”, 日本音響学会春季研究発表会講演論文集, pp. 271-272, 2007.

[16] P. A. Taylor, A. W. Black and R. J. Caley, "The architecture of the Festival speech synthesis system", Third international workshop on speech synthesis, pp. 147-151, 1998.

[17] J. Kominek and, A. W. Black, “CMU ARCTIC database for speech synthesis ver. 0.95“, Language technologies institute, Carnegie mellon university, 2003.

[18] VMware, Inc., VMware homepage, Retrieved from http://www.vmware.com/

[19] Red hat, Inc., Cygwin Homepage, Retrieved from http://cygwin.com/