ユーザーに合わせた対話ペースの調節

岩瀬 竜也*
東京大学工学部 械情報工学科 算法設計工学研究室

概要

本研究ではコーパス分析に基づき、音声認識を使わず、韻律情報を使って、人に合わせた自然なペースで道案内をするシステムを作った。そしてそのシステムを実験し、評価した。その結果、コンピューターだと気づかないほどの会話の自然さを達成することができた。また、音声認識システムの欠点を補うための韻律情報の可能性を考察した。

1 はじめに

現在、さまざまなタスクのもとで人間と対話するシステムが作られている。そして現在の対話システムは、音声認識を使ってユーザーの発話の言語情報を引きだし、それを使ってタスクを処理するものが多い。そして簡単なタスクを達成することは現時点ではほぼ達成していると言ってよい。しかし、これらのシステムと実際に会話してみると、人間と話している時のような自然さが感じられない。コンピューターと話しているのだと思い知らされ、人間と会話する時のような気軽さがない。その不自然さは具体的に言うと次の通りである。

- 認識時間が長く、人間の発言に対する反応が遅い。
- はっきり正確に発音しないと聞きとってくれない。登録していない語彙は処理してくれない。
- ユーザーの声の抑揚に合わせた、適切なフィードバック（あいづちなど）ができない。

そこで本研究では、計算が速く、人の声の抑揚を表現する韻律情報を用いて、音声認識の欠点を解消し、人間と自然に会話をするシステムを作った。具体的に言うと、本研究で追求した自然さとは次の四つである。

1. 相手の発話に対する十分に速い応答時間
2. 相手の発言への適切なフィードバック
3. 意味的におかしくない適切な会話内容
4. 適切な発言のタイミング

また、タスクは道順案内で、ユーザーはシステムのいう道順を聞きながらメモするというものである。

*tatsuya@sanpo.t.u-tokyo.ac.jp

案内人：これは新宿通りですな。
聞き手：　　新宿通り…はい。
案内人：でまっすぐ行って今度四谷四丁目左折。　　うん。
聞き手：　　左折ね？　　はい。
案内人：これが外苑西通り。
聞き手：　　外苑西…通り、はい。

図 1: 道順案内会話の例

本研究の目的はユーザーが気楽に話せる、自然な道順案内をするシステムを韻律情報のみを使って作ることである。さらに本研究では会話の自然さを評価する方法を考案した。そして実験により本システムを評価した。

2 関連研究

Schmandt[1] は本研究と同じく、韻律情報のみを使って人に道順案内するシステムを作った。しかし、Schmandt の場合は会話の自然さについての研究ではなく、道順案内というタスクを達成することを目的とした研究であった。これに対し本研究では、タスク達成よりも自然な会話を達成することを目的とする。

Ward、塚原 [2] は、一般の会話中に起こる「あいづち」の発生するタイミングは、韻律情報を使って予測できることを述べた。そして、話し手の 700ms 以上の発話において、低ピッチが 110ms 以上続いたら、低ピッチ開始時点から 350ms 後に聞き手のあいづちが打たれる、というあいづち予測法則を作った。そしてその法則をコーパスについてシミュレートし、再現率、適合率という 2 つの点数で評価した。本研究の目的の一つに、ユーザーの発話に対する適切なフィードバックを出す、ということがある。本研究では人間の会話中にもっとも頻繁に出されるという理由からフィードバックとしてあいづちを選んだ。そしてそれを出すタイミングを決める際にこのあいづち法則を参考にした。

3 コーパス分析

節 1で述べた 4 つの自然さのうち、会話の意味的内容の自然さが達成されるか否かは、その会話のモデルがきちんと作れるかどうかにかかっている。そこで、道順案内会話のモデルを作るため、人間同士の道順案内会話のコーパス分析をした。コーパスは 2 人の道順案内会話を

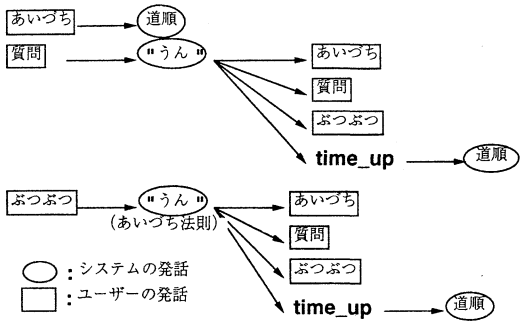

図 2: 会話モデル

１２人から１０個とった。会話をする状況としては、道順を教える側は地図帳を見ながら道順を言い、聞く側はそれをメモしながら聞く。その道は知らない。電話で話すように、互いに顔を見合わせずに行なった（図 1）。以下にその分析結果を示す。

発話	回数	割合
あいづち	562 回	80.4 %
ぶつぶつ	63 回	9.0 %
質問	37 回	5.3 %
その他	37 回	5.3 %

聞き手の発話回数

まずこの聞き手の発話回数のカウントから分かることは、自然な道順案内会話を再現するモデルを作る場合、聞き手の発言は、あいづち、ぶつぶつ、質問の３つだけをとりあえず考えれば良い、ということである。ここでぶつぶつというのは、聞き手が道順をメモする時に、その道順をぶつぶついいながら書きとる時の声のことである。

次に道順案内会話において、ある発話の直後にはどういった内容の発話が続きやすいか、発話の接続傾向を調べた。

案内人の道順	81.0 %
よそごと	9.3 %
その他	9.7 %

聞き手のあいづちの後

案内人のあいづち	50.7 %
案内人の解答	40.6 %
よそごと	4.3 %
その他	4.3 %

聞き手の質問の後

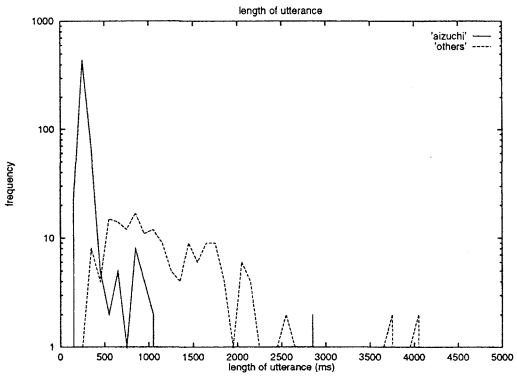

図 3: 聞き手の発話長さ (—: あいづち、- - -: その他)

これからわかることは、聞き手のあいづちは「了解」の意味があり、それが出されると案内人は道案内を先へ進める、ということと、聞き手が質問をした後には案内人が回答する、ということである。また、聞き手のぶつぶつの後の接続も調べたが、特に次に続きやすい発話はなかった。

また、分析していて、聞き手のぶつぶつの後に案内人があいづちを打つ場合と打たない場合があることがわかったが、この聞き手のぶつぶつの後の案内人のあいづちはあいづち法則 [2] に従うかどうか調べた。すべてのぶつぶつについて、あいづち法則の予測あいづちを見て、再現率、適合率を計算した。予測あいづち点から 500ms 以内に人間の出したあいづちが入っていれば、その予測は「当たり」とする。この時、再現率（当たりの数／人間が出した数）は 58.8 %、適合率（当たりの数／予測の数）は 47.6 %であった。ランダムに人間と同じ頻度であいづちを打った場合（27 %、22 %）と比べると、あいづち法則は有効であることがわかる。

以上を踏まえて道順案内会話のモデルを作ると図 2のようになる。なお、図中の timeup の時間は、臨時的に聞きての発話直後から、案内人が道順を出すまでの時間の平均値の２秒に設定してある。

また、ユーザーの発話をあいづち、ぶつぶつ、質問の３つに分類する基準を決めるため、コーパスを分析した。

図 3より、あいづちは発話長さ 500ms 以下で検出できることがわかる。

また、ぶつぶつと質問のピッチを調べたところ、表 1の通りになった。ピッチの勾配の計算には、発話直後 200ms 区間のピッチの値の最小２乗法による回帰直線の勾配を使った。表 1より、 yes/no 疑問文は wh 疑問文やぶつぶつに比べピッチ上昇の割合が高いことがわかる。よって文末のピッチ上昇により yes/no 疑問文を検出する。

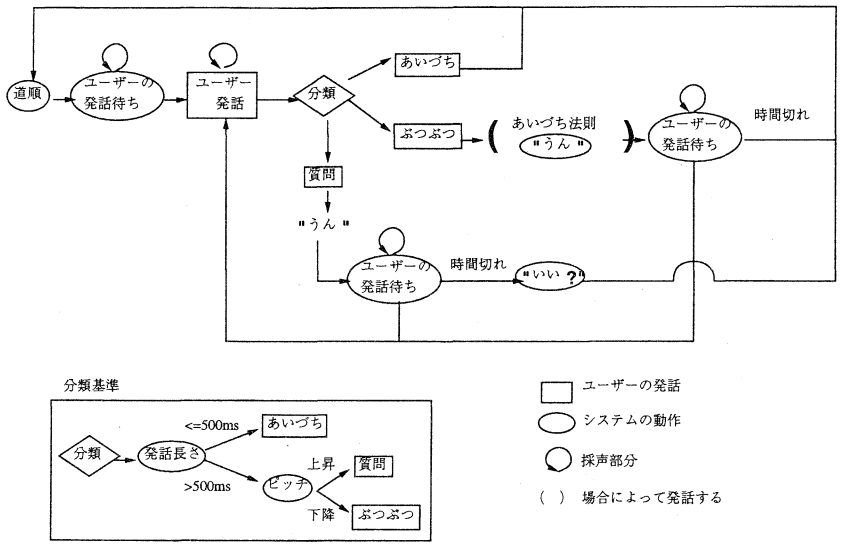

図 4: システム動作図

文の種類	文末のピッチの上昇の割合
yes/no 疑問	44%
wh 疑問	31%
ぶつぶつ	25%

表 1: 疑問文とぶつぶつの文末のピッチ

4 システムの構成

システムの動作は図4の通りである。このシステムは、節1で述べた4つの自然さを実現するように設計されている。

1. ユーザーの発言に対し、韻律情報のみを使って反応するため、音声認識と比べ十分に速い反応時間が得られる。2. ユーザーのぶつぶつに対して、あいづち法則 [2] を使ってあいづちを打つため、ユーザーの声の抑揚に合わせた自然なフィードバックが実現される。3. 会話の進め方はコーパス分析によって作られた会話モデルに基づいているので、会話内容の自然さも実現される。4. ユーザーが話している間はシステムは次の道順は出さず、ユーザーが了解の合図であるあいづちを出した時にシステムは道順を話すので、発言のタイミングは自然に調節される。

5 評価方法

本研究では、韻律情報を使って人と自然な会話をするシステムを評価する方法を考案した。そのようなシステムを評価する際に調べるべき点は次の2つである。

- 会話の自然さ
- 音声認識の欠点を補うための韻律情報の可能性とその限界

これらのことを調べるために、本研究では、実験でユーザーにシステムと会話してもらった時のアンケート、研究のことを一切知らない第3者に実験会話のテープを聞かせた時のアンケート、実験会話のコーパス、の3つを分析した。

ユーザーのアンケート、第3者のアンケートは会話の自然さ、スピードについてのいくつかの質問である。また、実験会話のコーパスは、ペース調節に失敗した会話、ユーザーの書いた道順のメモ、ユーザーの発話の分類正解率、ユーザーの「待って」、「それから？」という発言の回数、などを分析した。

6 実験、評価

日本人男性10人の被験者に、実際にシステムを使ってもらった。そして道順案内が終った後にアンケートに答えてもらった。また、別の10人の第3者に実験会話のテープを聞いてもらい、それに関するアンケートに答えてもらった。

まず、録音された実験の会話を聞いてみたところ、会話の各時点において会話のペース調節がうまくいかない時はいくつかの決まった理由があるということが分かった。このペース調節がうまくいかないパターンは次の4つであった。

1. （落あいづち）了解の合図である被験者のあいづちを正しく認識しなかったため、次の道順をすぐには出さず、道順案内が遅くなっている。

2. （落黙り）被験者が道順のメモをとっている間の無

声区間をノイズ（被験者の息など）の影響によって正しく認識せず、次の道順を出してしまい、道順案内が速くなっている。

3. （過あいづち）被験者のぶつぶつや質問などの発言を、了解の合図であるあいづちと間違って認識してしまい、次の道順を出してしまったため、道順案内が速くなっている。

4. （過黙り）被験者が無口な人、あるいは緊張しているような場合、あまり喋ろうとせず、そのためシステムは次の道順を出さず、道順案内が遅くなっている。

各実験会話のこのペーシングパターンを、「道順案内のスピードについてどう思いますか？」というユーザー、第3者へのアンケートと比べてみた。すると、速いパターンや遅いパターンが出ている会話については、アンケートの答えも速い、遅い、となっているのに対し、特にパターンが出ていない会話についてはアンケートの答えも普通、となっていた。このことは、システムがユーザーの発言を誤認識することなく正常に働けば会話のペースがうまく調節されることを意味する。これより適切な発言のタイミングは達成されたことがわかる。

その他のアンケート結果は、まず「会話は変でしたか」の質問に対してユーザーの80％、第3者の70％が普通と答えた。また、第3者にはコンピューターが道順案内していることは伏せてテープを聞かせるが、「実は道順案内していたのはコンピューターです」と教えたところ、90％が「驚いた」と答えた。これらのことより、本研究により会話の自然さは達成できたことがわかる。しかしながら、第3者に人間同士の道順案内会話を聞いてもらい、先の実験会話と比べてもらったところ、70％が人間同士の会話の方が自然に聞こえる、と答えた。その理由を聞いたところ、人間同士の会話の方が、

- 会話のテンポが速い
- 語彙が多い
- 言い間違い、「えーと」が多い
- 相手の発言の途中でも喋り出す
- キーワードの後など、あいづちのタイミングがつぼを得ている

ということが挙げられた。これらは本システムに足りない「人らしさ」として今後の課題になる。

また、ユーザーのアンケート、第3者のアンケートともに、「コンピューターとは話しにくい、話しにくそう」という意見が多かった。実験の前に「システムは自然に会話できるので、普通に会話して下さい。」と説明しても、意識してしまうユーザーが多かった。このことより、会話の自然さを評価する実験を行なう時には、話す相手がコンピューターだということを隠しておこなう必要があることがわかる。

また、ユーザーの70％は正しくメモがとれており、タスク達成率は70％となっている。また、ユーザーの発話のあいづち、ぶつぶつ、質問の分類成功率は58.9％、そのうちあいづちのみの抽出成功率は85.3％であり、あいづちに関しては高い分類精度が得られた。「信号を右ですか？」という質問に対する反応は、音声認識 (Sparc Station 20, HTK2.0) が4秒かかるのに対して、本システムは1秒もかからなかった。以上のことより、韻律情報は音声認識の欠点を補うのに有効であることがわかる。

7 おわりに

本研究では、言語情報を使わず、韻律情報だけを使って人に合わせた自然なペースで道順案内するシステムを作った。また、韻律情報を使って人と自然な会話をするシステムを評価する方法を考案した。

本研究により、第3者が聞いてもコンピューターとの会話だと気づかないほどの会話の自然さが達成された。

本研究により、対話のペース調節では、ユーザーが喋っているか黙っているかをきちんと認識することと、ユーザーのあいづちをきちんと認識することが重要であるということがわかった。また、本研究のように少ない語彙のタスクにおいては、韻律情報だけで会話が可能であるということがわかった。

また、本研究では達成できなかった会話の人間らしさとして、会話のテンポの速さ、語彙の多さがある。この人間らしさは音声認識と韻律情報を一緒に使用することにより達成できると思われる。

また、現時点においてはコンピューターとは話しにくい、という人が多いため、今後人と自然に会話するシステムを実験によって評価する場合、相手がコンピューターであるということを隠して実験する必要がある。

8 謝辞

本研究は、財団法人中山隼雄科学技術文化財団の資金援助を受け行なわれました。感謝致します。また、研究室のNigel Ward助教授には、研究の基本方針への助言を頂き、大変感謝しております。

参考文献

[1] Christopher Schmandt : "*Voice Communication with Computers*", VNR Computer Library (1994) pp.199 – 204.

[2] Nigel Ward, Wataru Tsukahara : "*Production of Back-Channel Feedback in Japanese may involve a Prosodically Triggered Reflex*", submitted to Language (1998)