応答音声から本人を確認する方式

2J－5

野田 晴義, 阪本 正治

日本アイ・ビー・エム (株)　東京基礎研究所

1　はじめに

近年、電話を利用したテレホンバンキングやテレホンショッピングなどのサービスが急速に拡大しつつある。これらのサービスに対して、話者照合技術 [1, 2] で音声による本人確認ができれば、利用範囲が大きく拡大していくと考えられる。

本稿では、利用者の負担を軽減するため、利用者とシステム (オペレータや音声自動応答システム) との間で交わされる応答音声を蓄積しておき、話者照合に利用する方式を提案する。

2　考慮点

2.1　話者照合とは

話者照合とは、申告された利用者の音声が本人のものであるかを判定する技術である。利用者の発声から特徴量を抽出し、学習データとして話者のHMMによる統計的なモデルを作成する。照合では入力された音声から申告者のモデルに対しての尤度 (確率) を計算して、閾値と比較する。また、照合はテキスト依存型、独立型、指定型に分類することができる。

2.2　信頼性

利用者が発声した声を録音して悪用することを防ぐためには、キーワードを可変にするなどして、本人が指定されたキーワードを正しく発音したときだけ受理される、テキスト指定型のような発声内容を確認するメカニズムが必要である。

2.3　発声の時間

話者のモデル化の統計的な精度を上げるために、学習において、照合に必要とされる音韻を含んだ発声データをできるだけたくさん与える必要がある。

しかしながら、電話による利用では、システムとの通話時間は短く、応答で発声するそれぞれの語句の長さも1～3秒と短いものが多い。また、照合用の語句を利用者が暗記しておいたり、システムからその都度与えられた語句を復唱することも難しいことがある。

2.4　発声内容

発声データが難しいもしくは長いと、正確に記憶していることは難しい。そのため、正確に発声するためには事前に用意した印刷物を見なければならない場合が多い。

また、個室や電話ボックス以外では、発声の内容を他人に聞かれてしまう可能性があるため、固定パスワードなどのように機密性の高い内容をそのまま利用することはできない。

3　応答音声の利用

これらの問題点に関して、実際の運用においての解決策を検討した。我々が想定したアプリケーションとは、電話によるサービスをシステムが行なっており、利用者との間で短い応答が頻繁に行われる場合である (図1)。

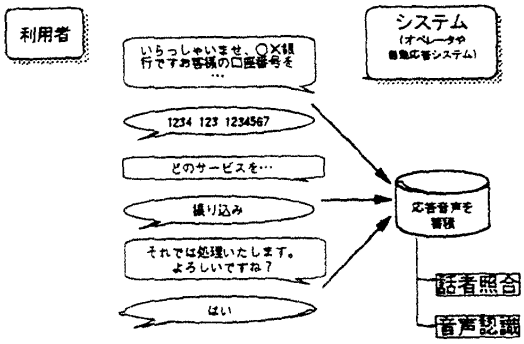

図 1: 適用例

3.1　発声データ量と照合率

学習や照合において、発声データが少ないと安定した話者の特徴を得ることができない。そこで我々は、学習と照合において発声データ量が照合率に及ぼす影響を実験した。実験データには、数字に限定したYOHOコーパス [3] を用いた (図2)。

その結果、15～20秒の学習用の音声データがあれば、照合用のデータが約2秒の時にも90%を越える照合率が得られるという知見を得た。

3.2　話者照合と音声認識の融合

システムからのプロンプトに対する利用者の返答は固定、複数からの選択、可変なものがある。我々が想定しているアプリケーションでは、システムからのプロンプトが与えられた時点では返答の内容が一意に決まらない場合が多い。そのため、テキスト指定型 [4] のように照合時に指定したテキストに基づいて固定の音響モデルを連結する方式ではすべての応答に対処できない。

Speaker verification by using responding utterances
Haruyoshi Noda, Masaharu Sakamoto
IBM Research, Tokyo Research Laboratory

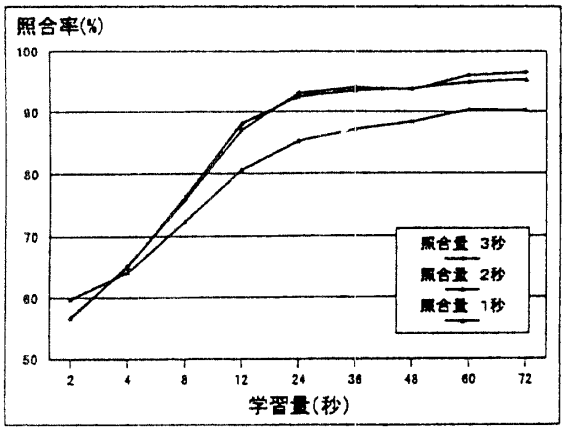

図 2: 発声量と照合率 (YOHO)

そこで、テキスト独立型の話者照合と音声認識を組み合わせる方式を考えた。この方式では、発声内容を音声認識により確認する機能と発声を照合して本人確認をする機能を組み合わせることにより、信頼性の高さを保ちながら発声の内容に自由度を与えられるものである。

4 実験

4.1 実験システム

我々の実験では、次のようにして前述の機能を実現している。各発声データから音声区間を検出し、12 次元のメル・ケプストラム、パワー、Δケプストラム、Δパワーを特徴量として抽出する。その特徴ベクトル時系列を 1 状態 32 混合ガウス分布で話者と特定話者の単語のモデルをそれぞれ作成する。照合では、同様にして得られた特徴量から話者のモデルに対して求められた尤度を正規化して判定を行う。閾値は各話者毎に 2 種類の誤り (詐称者受理率、本人棄却率) が等しくなる値（等誤り率）に事後的に設定した。

回線の特性に対してはケプスルトラム平均正規化、回線や背景の定常的な雑音に対してはスペクトル・サブトラクションを用いている。

4.2 実験方法

システムの中で使用される応答のうち、照合用セットとしては、数字、名前、住所、コマンドの一部を選択した。今回の実験では、照合のみに応答音声の利用を行なった。学習は、上記の照合用語彙セットを数回繰り返した。

1. 利用者は自分の話者番号を申告する
2. システムからのプロンプトに対して、音声で返答をする
3. システムが音声データを認識して、次の処理へ進む
4. 認識された単語が照合の語彙セットに含まれていれば、照合用に蓄積される
5. 蓄積量が一定量に達すると自動的に照合が行なわれ、本人確認の判定が行なわれる

4.3 実験結果

1. 短い応答音声を用いて、93～97%の照合率が得られた
2. 応答音声を利用することにより、次のメリットが明らかになった。
   サービス提供者：利用者に気づかせずに照合が行なえる。音声認識と組合せることにより信頼性が高められる。
   利用者：照合用の発声が不要。

5 まとめ

本稿では、利用者の照合に対する負担を軽減するために、システムと利用者の間で交わされる応答音声を蓄積し、効率的に活用する方式を提案した。そして、実験により実用的な照合率を得ることが確認できた。

今後の課題としては、この方式を照合だけでなく、学習の時点において適用する実験を行なっていきたい。

参考文献

[1] 古井：「ディジタル音声処理」, 東海大学出版会.
[2] 松井：「HMM による話者認識」, 電子情報通信学会技法, SP95-111, 1995.
[3] Jpseph P. Campbell, Jr, “Testing with the YOHO CD-ROM voice verification corpus”, IEEE, 1995.
[4] 松井：「テキスト指定型話者認識」, 電子情報通信学会論文誌, D-II, Vol.J79, No.5, 1996.