大規模録音音声データーベースにおける録音誤りの検出の検討*

村上仁一 （NTT 情報通信研究所） 鈴木博和 （NTT アドバンステクノロジ)

1 まえがき

本論文では、音声認識を利用して録音音声データベースの誤りを検出する方法の提案と、その検出結果について述べる。

大規模な録音音声データベースを作成する場合、大量のラベル（読み）と音声データが対になったファイルを作る必要がある。しかし、これらは手作業であるため、ラベルと音声データが一致しないファイルが出現する。この誤りを削減するに、人手による再検聴を繰り返す必要がある。本論文では、検聴をするファイルを減らすために、音声認識装置を利用した方法について述べる。そして実際データベースを作成したときの結果について報告する。

2 録音音声データベースの作成

2.1 録音音声データベースの作成方法

録音音声データベースを作成するには複数の方法がある。しかしコストの面から大規模なデータベースは以下の方法で作成されている。

1. リスト作成部

   発話リスト（録音リスト）を作成する。

2. 収録部

   ナレータにスタジオで発話リストを発声してもらう。この音声を DAT に収録する。

3. 音声切出部

   DAT で収録された音声を、発話リストごとに切り出す。（以下これを音声データと呼ぶ。）

4. ラベル付与部

   音声データに対応するラベル（読み）を作成する。

5. 検聴部

   ラベルと音声データを人間が検聴する。（以後、この２つのデータをファイルと呼ぶ。）誤って収録された音声データは、再度ナレータによって録音する。

2.2 問題点

上記のように録音音声データベースを作成した場合、5 の検聴部において、人間が多くのファイルを検聴するため、ラベルと音声データの発話内容が異なるファイルを見過ごしてしまう可能性がある。この誤りをなくすためには、全ファイルに対し人間による検聴（再検聴）を繰り返す必要がある。

3 音声認識を利用した誤り検出方法

3.1 単語音声認識を利用した誤り検出方法

本論文では、次の仮説を立てることで人間が再検聴するファイルの数の削減を行なった。

> 仮説１：単語音声認識結果とラベルが一致しているファイルは、ラベルと音声データが一致しているとみなす。

この仮説に従い、ラベルと音声データが一致したファイルの人手による再検聴は行なわない。しかし、大規模な録音データベースでは認識語彙数は数十万件になる。そのため、多くのファイルが誤認識されるため再検聴の数の削減の効果は少ない。そこで次の仮説を加えた。

> 仮説２：ラベルと録音データの発話内容が一致しない誤りの多くは、発話リストの前後のラベルを発話している。

この仮説に従って、単語認識の語彙は、再検聴する音声データの発話リストの前後 N 件のラベルとした。したがって認識語彙数は $N+1$ 単語になる。

図１に $N=2$ のときの認識語彙を示す。この図ではファイル番号 4 の音声データを再検聴するときの様子を示している。単語認識の語彙は「大阪、埼玉、名古屋、静岡、神戸」の 5 件である。そして「名

---

*“ A Study of Speech Data-Base Error Detection ” by Jin’ichi Murakami (NTT Information and Communication System Laboratories) and Hirokazu Suzuki (NTT Advanced Technology)

古屋」以外が認識された場合、音声データは誤っている可能性があると見なして、人手による再検聴を行なう。

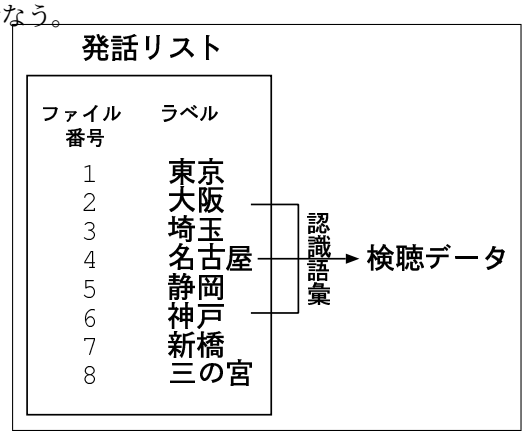

図 1: 単語認識装置による誤りの検出

3.2 フローチャート

本論文で提案する再検聴のフローチャートを以下に示す。

1. 音声データを単語認識装置で認識する。認識語彙は再検聴するラベルの前後 N 件とする。
2. 認識結果とラベルを比較する。
3. ラベルと認識結果が一致しないファイルを出力する。
4. 出力されたファイルを人間が再検聴し、実際にラベルと音声データが異なるファイルを選出する。
5. ラベルと音声データが異なる音声データを、再度録音する。

4 実験

4.1 DB 作成

PB 電話番号案内実験システムは、PB 電話器を用いて音声ガイダンスを聞いて電話番号案内をするサービスである [2]。このシステムのために、ナレータ 6 名によって、住所・姓名・企業、合計 60 万件を録音した。このデータベースを人手で 1 度検聴したあと、本論文で提案した方法で誤りの検出を行なった。

4.2 単語音声認識

単語音声認識には、HTK[1] を利用した。音素モデルは、ATR の C set 女性話者 32 名, 1600 文から不特定話者モデルを作り、次に話者ごとに 100 単語の連結学習をして HMM のモデルを作成した。分析パラメータの条件を表 1 に示す。

4.3 結果

実験では $N = 20$ とした。結果を表 2 に示す。

表 1: 音素モデルの学習条件

音響モデル	4 状態 3 ループ混合分布型 HMM
混合数	10 混合 full covariance
音響パラメータ	log power + 12 次 FFT melcep + Δ log power + 12 次 Δ FFT melcep
フレーム長	5ms
フレーム窓長	25ms
sampling 周波数	16KHz

表 2: 実験結果

録音対象	録音件数	検出件数 (%)	誤りの件数 (%)
住所	193881	9021 (4.65%)	103 (0.053%)
姓名	186547	15596 (8.36%)	34 (0.018%)
企業	109093	3273 (3.00%)	134 (0.122%)
企業 2	70683	2209 (3.13%)	15 (0.021%)

これらの結果から、単語認識において平均約 5.3% が認識結果とラベルが一致しなかった。これらのファイルを人手によって再検聴したところ、平均 0.05% は誤っていたことがわかった。

5 考察

今回の録音データベースで使用された発話リストはソートされている。そのため発話の良く似たラベルが続いている。このため検出精度が高くなったと思われる (5%)。発話リストをランダムにして作成した場合、この値は低くなると思われる。

6 まとめ

本論文では、音声認識を利用した録音音声データベースの誤りを検出する方法の提案と、その実験結果について述べた。この結果、全データの 5%が検出され、これらのファイルを人手によって再検聴したところ平均 0.05%は誤っていた。これによって本手法の有効性が示された。

参考文献

[1] Cambridge University Engineering Department Speech Group and Entopic Research Lab Inc., “HTK:Hidden Markov Model Toolkit V1.5” (Sep. 1993).

[2] 東田 他, “オペレータレス自動電話番号検査システムの開発”, 自然言語処理研究会 98-NL-123-4 pp.25-32 (Jan. 1998).