平成 11 年 2 月 23 日

電子・情報工学専攻	学籍番号	947455
申請者氏名	Markov Konstantin	

指導教官氏名	中川 聖一 梅村 恭司 奥山 徹

論 文 要 旨 (博士)

論文題目	Text-Independent Speaker Recognition Based on Frame Level Likelihood Transformations (フレームレベルの尤度変換に基づくテキスト独立型話者認識)

(要旨 1,200 字程度)

電子計算機と音声処理技術の発展により、コンピュータによって人間の音声を認識することが可能になってきた。更に、コンピュータによって話している人を同定することも可能になってきた。しかしながら、話者認識技術を広くアプリケーションに適用するためには、なお解決されるべき多くの問題がある。話者認識システムは、顧客要求サービスのセキュリティに対して大きな潜在的需要を持っている。これらのサービスは、電話を通して行なわれるバンキングトランザクション、データベースからの個人情報アクセス等を含んでいる。

本研究では、ガウス混合モデルに基づいた、テキスト独立話者認識に関する考察を述べている。いくつかの新しいアプローチを提案し、議論し、評価している。それらはすべて、話者識別への各短時間毎に得られる尤度の貢献を評価するためフレーム毎の尤度をある適切に選んだ関数によってスコアに変換するフレームレベル尤度変換技術に基づいている。そのようなある種の関数は尤度正規化を実行する。フレーム尤度正規化は対象話者以外の話者をモデル化したバックグラウンドモデルを使用して行なわれる。バックグラウンドモデルの中で最も良かったものは、音響的に近接している話者集合のモデルからなるコホートセットと呼ばれるものである。我々が開発した他のタイプの変換関数は、尤度を他のモデルの尤度との関係を考慮したスコアに変換する。変換関数で用いる重みは、トレーニングデータ上のシステム性能の統計分析から決定される。我々がランク重み付け（WMR）と呼ぶこの技術は、フレーム尤度正規化よりさらに有効であることが分かった。

評価実験は、両技術が話者の声質の変化や話速の変化に対して頑健なことを示した。特に、WMR 技術で、5 時期にわたって集められた 35 名の話者の音声からなる NTT データベースで 99.1%の話者識別精度を達成した。有名な TIMIT データベースでは、630 の話者に対して 99.6%の精度を得た。一方、話者照合等価誤り率は、NTT と TIMIT のデータベースに対してそれぞれ 0.12%および 0.01%であった。

次に、フレームレベル尤度正規化概念をガウス混合分布の学習まで拡張することによって、最大正規化尤度 (MNL) 推定アルゴリズムと呼ばれる新しい識別学習方法を開発した。その名前が示唆するように、学習データのフレームレベルで正規化された尤度を最大化するアルゴリズムのことである。正規化された尤度を最大化することは、ターゲットクラスと正規化のためのバックグラウンドセットの役割をする他クラスとの間のよりよい分離を導く。この場合、ある話者が類似した音響的な特徴を持っている人に誤って分類されることが頻繁に起こるので、コホート型のバックグラウンドセットが最も適している。他の識別学習方法とは対照的に、MNL 目的関数の最適化のために EM アルゴリズムを利用した。この時、解へ収束を保証するために、繰り返し再推定式のためのアルゴリズムのいくつかの修正を行なった。MNL 学習アルゴリズムは、パターン認識分野で有名な最小分類誤り学習 (MCE) 方法より良い性能を示した。特に、雑音を含んだ電話音声データに対しては、著しい改善を示した。

最後に、異なる話者情報源の統合法を考察した。LPC ケプストラム係数は、短時間音声スペクトルの表現として広く使用される。しかしながら、いくつかの研究が示すように、予測誤差信号（LPC 残差）は話者情報を含んでいる。LPC ケプストラム係数でのスペクトル情報と LPC 残差から抽出された情報、さらに音声の基本周波数やピッチに含まれる話者依存の情報を統合する話者認識システムを開発した。このシステムを構築する際に注意を払った点は、異なった情報源間の相関を考慮したことである。評価実験では、ピッチとケプストラム係数の間の相関を考慮した場合が最も有効なことが示された。