氏　　名	山田　誠
学位（専攻分野）	博士（統計科学）
学位記番号	総研大甲第1333号
学位授与の日付	平成２２年３月２４日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第６条第１項該当
学位論文題目	Kernel Methods and Frequency Domain Independent Component Analysis for Robust Speaker Identification
論文審査委員	主　査　教授　福水　健次 准教授　池田　思朗 准教授　杉山　将（東京工業大学）

The speaker identification is one of the key technologies for person identification in humanoid robots. Especially, when the face information is not available, the speaker identification is the only way to identify person, thus, to improve the speaker identification performance is an important issue for person identification tasks. This doctorial thesis concerns four major issues in speaker identification for humanoid robots in practice, and proposes solutions, which can be implemented in speaker identification systems.

First, the humanoid robots should identify the speaker in real-time with high identification rates. In these days, the kernel methods such as the support vector machine (SVM) and kernel logistic regression (KLR) are popular for speaker identification tasks, and the kernel based systems outperform the conventional Gaussian Mixture Model (GMM) based system. However, the kernel based speaker identification systems are usually computationally intensive, and this is of course not preferable for real-time implementation. To deal with the computational issue, the thesis proposes a method of approximating the sequence kernel that is shown to be computationally very efficient in Chapter 4. More specifically, the thesis formulates the problem of approximating the sequence kernel as the problem of obtaining a *pre-image* in a reproducing kernel Hilbert space. The effectiveness of the proposed approximation is demonstrated in text-independent speaker identification experiments with 10 male speakers; the proposed approach provides significant reduction in computation time while performance degradation is kept moderately. Based on the proposed method, the thesis develops a real-time kernel-based speaker identification system using the Virtual Studio Technology (VST).

Second, the speech features vary over time due to session dependent variation, the recording environment change, and physical conditions/emotions. However, conventional kernel based systems implicitly ignore these facts, and they just simply assume that the training and test input probability distributions of the training and test datasets are same at any time. To alleviate the influence of session dependent variation, it is popular to use several sessions of speaker utterance samples or to use *cepstral mean normalization* (CMN). However, gathering several sessions of speaker utterance data and assigning the speaker ID to the collected data are expensive both in time and cost and therefore not realistic in practice. Moreover, it is not possible to perfectly remove the session dependent variation by CMN alone. Thus, in Chapter 5, the thesis proposes a novel semi-supervised speaker identification method that can alleviate the influence of non-stationarity such as session dependent variation, the recording environment

change, and physical conditions/emotions. The thesis assumes that the voice quality variants follow the *covariate shift* model, where only the voice feature distribution changes in the training and test phases. The proposed method consists of weighted versions of kernel logistic regression and cross validation and is theoretically shown to have the capability of alleviating the influence of covariate shift, where the weight (a.k.a importance) is estimated from the training and test distribution using the Kullback-Leibler Importance Estimation Procedure (KLIEP). The thesis experimentally shows through text-independent/dependent speaker identification simulations that the proposed method is promising in dealing with variations in voice quality.

Third, the humanoid robots are desired to automatically detect the unknown speakers information into the dictionary, and the speaker detection task can be formulated as the outlier detection problem (i.e., outliers can be the unknown speakers). Since the outlier detection problem can be solved through the comparison between the log likelihoods of the unknown speaker and the speakers, the estimation accuracy of the log likelihoods is an important issue to improve the speaker detection performance. Thus, in Chapter 6, the thesis propose a new importance (a.k.a likelihood) estimation method using Gaussian mixture models (GMMs) and principal component analyzers (PPCAs) mixture, where the proposed approach estimates the importance without going through the density estimation. An advantage of the proposed methods is that covariance matrices or projection matrices can also be learned through an expectation-maximization procedure, so the proposed method expected to work well when the true importance function has high correlation. Through experiments of outlier detection, the thesis shows the validity of the proposed approaches.

Forth, the humanoid robots move throughout the world, and the surrounding environment, source positions, and source mixtures are constantly changing. In addition, the speech overlaps are frequently occurred during conversation. Thus, the source separation techniques are useful for improving the speaker identification performance. To deal with those problems, in Chapter 7, the thesis considers the problem of two-source signal separation from a two-microphone array, where a point source such as a speech signal is placed in front of a two-microphone array, while no information is available about another *interference* signal. The thesis proposes a simple and computationally efficient method for estimating the geometry and source type (a point or diffuse) of the interference signal, which allows us to adaptively choose a suitable unmixing matrix initialization scheme. The proposed method, *noise adaptive optimization of matrix initialization* (NAOMI), is shown to be effective through source separation and speaker identification simulations.

博士論文の審査結果の要旨

山田誠氏の博士論文の審査は，福水健次（主査，総合研究大学院大学），松井知子（総合研究大学院大学），池田思朗（総合研究大学院大学），杉山将（東京工業大学）の4名により構成される審査委員会によって，11月12日の予備審査委員会および1月26日の審査委員会を，本人および表記4名の委員全員の出席のもとに行った．その結果，全員一致で同氏の論文は博士（統計科学）を授与するに値するという結論に達した．論文の概要および評価は以下の通りである．

本論文は，音声による自動話者識別システムの要素技術に対する統計的手法に関する研究をまとめたものであり，全7章88頁からなる．まず1章では，研究の背景と本論文で議論する課題が述べられている．課題としてヒューマノイドロボット等に用いる自動話者識別システムを想定し，そのための要素技術として，実時間処理を目的とした音声特徴量計算の高速化，時間的環境的変動に対応するための共変量シフトの適用とその方法の改良，雑音除去技術としての独立成分分析の改良が述べられている．2章では，本論文の基礎となる技術が説明されている．本論文の主要結果を述べた3章から6章は個々の課題に対する方法を論じている．3章では，実時間での話者識別を目的として，可変長系列データに対して有効性が確認されている正定値カーネルMOSKの計算の高速化手法を提案し，実時間処理システムへの実装を行っている．4章は，各話者の音声の時間的環境的変動による識別率の低下に対処するため，変動の問題を共変量シフトによって定式化し，近年提案されたKullback-Leibler importance estimation procedure（KLIEP）という方法の話者識別への応用法を提案し，提案手法の有効性を確認している．また5章はKLIEPの改良法を提案し，その有効性を実験的に示している．6章は，話者以外の雑音を除去することを目的として，音源分離のための周波数領域での独立成分分析において，環境変化に対して適応的にパラメータ最適化を行う手法を提案し，話者識別の性能が向上することが示されている．7章は結論と今後の展開が述べられている．

本論文3章で論じられているように，実時間処理は実用的な話者識別システムにおいて必須の課題である．これに対し，識別率は高いが計算量が大きいMOSKカーネルの計算を高速化し，実時間処理が可能な統計的手法を提案しかつ実装まで行った点は応用研究として評価できる．また，4，5章で論じられた音声の時間的環境的変動は話者識別一般において重要な課題であるが，共変量シフトという統計的枠組みを適用する考えは話者識別課題においては新規性があり，実際に有意な性能改善を示した点は十分評価できる．6章において提案された雑音除去の方法は，指向性マイクロフォンやマイクロフォンアレイなど高コストな方法を用いることなく，指向性に関する適応的な雑音除去の新しい方法を提案しており，ヒューマノイドロボットなどの話者識別システムの要素技術として有効性が認められる．

また，論文誌への投稿状況として，4章の内容をまとめた論文が国際学術雑誌Signal Processingに採録決定している．また5章の内容を部分的にまとめた論文がIEICE Transactions on Information and Systemsに採録されている．3章の内容をまとめた

論文が査読付国際会議に採択済み，また 6 章の内容に関しては国際学術雑誌に投稿中である．

以上の理由に基づき，本審査委員会は，統計的手法の音声情報処理への応用を扱った論文として，本論文が博士（統計科学）を授与するに十分な内容を有するものと判定した．