様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２８　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2012～2014
課題番号：　２４６５００７９

研究課題名（和文）非示量性情報理論に基づく音声言語処理

研究課題名（英文）Spoken Language Proceeding Based on Non-Extensive Information Theory

研究代表者
篠田　浩一（SHINODA, KOICHI）

東京工業大学・情報理工学（系）研究科・教授

研究者番号：１０３４３０９７

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：音声言語処理に対し、従来の示量性統計理論を拡張した非示量性統計理論を適用する方法論を開発した。まず、音声認識のための特徴抽出について、周囲雑音・回線の違いから生じる変動に対し頑健な、q-log spectral mean subtraction (q-LMSN)手法を提案し、従来のCMNを用いた手法に比べ優位に性能が高いことを示した。また、音声認識・映像意味インデクシングにおいて、HMMやGMMの出力分布として、周囲雑音の変動に頑健なq-Gauss混合分布を用いる方式を提案し、その効果を確認した。

研究成果の概要（英文）：We have developed a methodology for spoken language processing based on non-extensive statistical theory, which is an extension from the conventional extensive statistical theory. We first developed q-log spectral subtraction (q-LMSN) to achieve robustness against the difference of environmental noises and of channels. We proved that it was significantly better than the conventional CMN. Next, we developed a recognition a method using q-Gaussian mixtures for output probabilities in GMMs and in HMMs. We applied it to speech recognition and to video semantic indexing and proved its effectiveness.

研究分野：　統計的パターン認識

キーワード：　音声情報処理　映像情報処理

1．研究開始当初の背景

現在の情報理論は、Boltzmann-Gibbs 統計とシャノン・エントロピーを軸とした示量性理論である。ここで示量性とは「系の大きさに比例する性質」のことで、特に相加性が成立することを意味する。この理論は、対象とする現象が独立事象の和として表現され、その和が保存されることを前提としている。従来、音声言語処理では、この示量性(extensive)統計理論を基盤としてきた。

音声言語処理は過去 20 年間で飛躍的進歩を遂げた。例えば音声認識では、静かな環境での丁寧な発声に対しては 90%以上の認識率である。そこでは、示量性の情報理論が基盤として大きく寄与している。一方で、雑音下の音声認識、会話音声の認識など、困難な問題が依然として多く存在している。音声言語は典型的な時系列パターンであり、長時間の相関をもつ事象が複数あり、それらが複雑に絡み合っている。また、指数型分布よりも、べき乗型分布によく従う現象がある。さらに、学習に用いる教師ラベルにノイズが多く含まれている。このような現象を独立事象の和に分解することは難しく、従って示量性理論に基づくモデルがうまく適合しないことが経験的に知られている。

一方で、近年統計物理学の分野で非示量性(non-extensive)統計理論の研究[1]が精力的に進められている。これは示量性統計理論を非指数的な体系に一般化するものであり、パラメータ q を導入した q-指数が用いられる。相加性は失われるものの、長時間相関の存在する時系列データなど複雑な事象のモデリングに特に有効であることが知られている。

示量性統計理論における一般化指数関数はパラメータ q をもつ。この関数は、q→1 のとき指数関数 exp(x)に漸近的に一致し、 q > 1 のときべき乗則を近似する。この理論における Tsallis エントロピーは同じパラメータ q をもち、q->1 のとき、シャノン・エントロピーに一致する。Tsallis エントロピーではもはや相加性は成立しない。すなわち、二つの系 A,B があるとき、

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B)$$

非示量性統計理論ではパラメータ q を適切に調節することにより、長時間相関をもつデータやべき乗則に従うデータを柔軟に表現可能である。例えば、経済学の各種指数は正規分布を仮定した場合の外れ値が予想よりも高頻度で発生することが知られる(リーマンショックなど)。それらの分布が、一般化指数を用いた q-正規分布でよく表現できることが最近分かってきた。他にも生物学・言語学などの多くの分野で、精力的に研究が進められている。しかしながら、現段階では、パラメータ q のもつ物理的意味は十分に明らかでなく、したがって、その値を求める理論も存在しない。

<引用文献>

[1] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," J. Stat. Phys. 52, 479 (1988).

2．研究の目的

非示量性統計理論を、音声言語処理における、(1)音声からの特徴抽出、(2)音声認識・話者認識のための音響モデル、(3)音声認識・機械翻訳における言語モデルの 3 つの課題に適用する。従来の示量性統計理論では対応が困難であった話者・周囲雑音・回線・タスクなどの外部条件の違いに対して頑健なモデル化手法を構築する。また、 q-指数のパラメータ q の最適化を通して、これらの違いに起因する現象の物理的特性を解析する。

音声言語は様々な時間スケールの事象が複雑に絡みあっており、個々の事象の物理的特性は十分に明らかでない。本研究は、これらの事象に対して、まず非示量性統計理論を用いてフィッティングを行い、次にそのパラメータの最適化を通してその物理的特性の解析を行う新しい方法論を提案する。

外部条件の違いに起因する性能低下に対し頑健であり、かつ、汎用性の高い音声言語処理技術となることが期待できる。また、物理的特性に関する解析で得られた知見は新しい音声言語処理フレームワークの創発につながる可能性がある。

また、非示量性理論は音声だけではなく、他のメディア、例えば、画像や映像に対しても有効であることが期待される。そこで、映像における意味インデクシングに対し、非示量性理論を応用し、その効果を評価する。

3．研究の方法

音声言語処理における 3 つの課題、(1)耐雑音音声処理、(2)音声認識・話者認識のための音響モデル、(3)音声認識・機械翻訳における言語モデルに対し、非示量性統計理論の適用を試みる。(1)に関しては回線・雑音、(2)については話者・発声スタイル、(3)については発声スタイル・タスクの違いに対して頑健なモデリングを開発する。既存のツールを改良してアルゴリズム実装を行う。主に既存の音声データベースを用いて手法の評価を行う。方式開発のため、高性能な計算サーバを購入し、利用する。

画像・映像処理においては、PASCAL VOC の画像データセット、映像検索のためのデータセットである TRECVID SIN タスクデータなどを用いて、非示量性理論の効果を検証する。

4．研究成果

(1) q-LMSN を用いた頑健な音声特徴量抽出

従来の音声認識は静かな環境での性能は極めて高いが、環境雑音の多い環境下ではその性能は著しく劣化する。これは音声特徴量が雑音の影響で変化し、静かな環境の音声を用いて作成された音声モデルとのミスマッチが生じているためである。環境雑音は加算性雑音と乗算性雑音に大きく分類される。ここで加算性雑音の例としては、路上騒音、計算機のファンの音、他の人の話し声などがあり、乗算性雑音の例としては、音の反響や回線歪みなどがあげられる。

従来の音声認識のための耐雑音処理では、音声特徴量を抽出する際に、周波数領域に変換後、音声と雑音との位相差を無視して対数スペクトルを抽出することで、これらの 2 種類の雑音を分離している。しかしながら、現実には位相の影響を無視できず、一定の割合で交差項(cross-term)が存在する。この交差項が従来手法での性能向上を妨げる一因となっている。

一方、今までの音声特徴量抽出は、示量性の統計理論に基づくものであり、この交差項も、その理論の枠組みの中で現れるものであった。通常の和の演算が保証されない、本研究では、非示量性の理論の枠組みを用いることで、交差項も含めた雑音重畳音声のモデリングを行うことが試みた。

従来、乗算性雑音の除去で有力な手段として、ケプストラム平均除去(Cepstral Mean Subtraction; CMS)があげられる。これは、音声特徴量であるケプストラムの長時間平均をとり、それを、ケプストラムから差し引くことで、対数スペクトル領域における減算、すなわち、スペクトル領域における除算を行う処理である。通常はこの CMS を適用する前に、スペクトル領域での減算(スペクトルサブトラクション)を行うことで、加算性雑音を除去しているが、前述の通り、完全に除去することは難しく交差項が存在することが課題であった。そこで提案手法では、非示量性理論を用いて定義された q-対数領域での減算処理を行うことで、加算性雑音と乗算性雑音の両方を効率よく除去する。ここで、q-対数領域は、(線形)スペクトル領域と、対数スペクトル領域の間に存在し、その中のどこに位置するかをパラメータ q で指定する。

提案手法を Aurola-II と CENSREC-2 の 2 つの雑音データベースで評価した。それぞれ、複数のマイクロフォンで収録された、複数の周囲雑音のデータと音声のデータがあり、音声と雑音を複数通りの SN 比(Signal-to-Noise Ratio)で音声に重畳したものである。その結果、いずれの場合でも、提案手法の単語認識率が、従来の CMS 法のそれを優位に上回る結果を得た。提案手法の有用性は明らかとなった。

今後の課題としては、制御パラメータ q の値をどのように決めるか、ということがあげられる。最適な q の値は、雑音の種類、SN 比のよって異なり、また、同じ雑音、同じ SN 比でも、周波数帯によって異なる。それら相違に対応した最適化手法を開発する必要がある。

(2) q-混合ガウス分布を用いた映像検索

近年, Youtube などをはじめとした、映像コンテンツ共有サイトの発展により、膨大な量の映像データが利用可能となった。現在の映像検索システムの多くは、映像の内容を表す意味的タグに基づいて映像を検索しているが、それらの意味的タグは人手で付与されており、自動的かつ高精度な意味的タグ付与技術は確立していない。

映像や画像データの意味的分類は、TRECVID ワークショップや Pascal Visual Object Classes Challenge で取り上げられており、映像中の物体やイベントの検出に関する様々な研究が行われている。その中でも、Bag-of-visual-words(BoW) 法は映像や画像の効果的な分類手法として注目されている。BoW 法は、SIFT に代表される局所特徴をベクトル量子化し、そのヒストグラムを新たな特徴量として識別学習の入力に用いる方法である。最近では、ベクトル量子化における量子化誤差を低減する方法として、混合ガウス分布(Gaussian mixture model; GMM) を用いる手法が提案されている。GMM は BoW 法を確率的枠組みに発展させた手法であると解釈することができ、BoW 法よりも高い映像・画像分類精度が報告されている。

ガウス分布は Boltzmann-Shannon エントロピーを最大化する確率分布として導出されるが、物理学における複雑系の分野では、Boltzmann-Shannon エントロピーを一般化した Tsallis エントロピーから導出される q-ガウス分布がマルチフラクタルなどのモデルの表現に効果的であることが知られている。q-ガウス分布は実数値をとるパラメータ q(q 値と呼ぶ)により、ガウス分布を拡張したものであり、q > 1.0 の時にガウス分布よりも裾の長い分布が得られ、q -> 1.0 でガウス分布に漸近する。q-ガウス分布では、q 値により裾の長さを変化させることで、2 次よりも高次なモーメントを調節できるため、外れ値に対する頑健性が向上することが期待される。このような背景のもと、q-ガウス分布の混合モデルである q-混合ガウス分布(q-GMM) を映像と画像のセマンティックインデクシングに適用し、その有用性をした。

評価実験では、PASCAL VOC 2010 (classification task, validationchallenge) と TRECVID 2010 (semantic indexingtask) のデータセットおよび評価基準を用いた。PASCALVOC 2010 データセットは各 5000 枚程度の学習用画像データとテスト用画像データから構成されており、20 種類の物体に関する意味的タグが付与されている。TRECVID2010 データセットはインターネットアーカイブから集められた 400 時間の映像データから構成されている。映像データにはショット境界情報(カメラの切り替わりフレーム番号)が付与されており、学習用の映像ショット数は 12 万程度、テスト用は 15 万程度である。また、意味的タグは物体、イベント、シーンに関する 30 種類が用意されている。評価尺度は各検出対象に関する Average Precision (AP) の算術平均である Mean Average Precision (Mean AP) を用いた。

局所特徴抽出では、SIFT 特徴(128 次元)と Hue histogram 特徴(36 次元) を連結した 164 次元の特徴量を、100x100 の格子点から 3 段階のスケールで抽出した。また、得られた特徴量の次元を主成分分析により 32 次元に圧縮した。TRECVID の映像データにおける局所特徴抽出では、映像ショット毎に指定されているキーフレーム画像からのみ特徴量の抽出を行った。q-GMM の混合数は 512、制御パラメータは 20.0 とした。また、識別器にはサポートベクト

ルマシンを用いた。

PASCAL VOC 2010 データセットにおける Mean AP は、通常の Bag-of-words 法を用いた場合 30.9%、q-GMM を用いたヒストグラムによる画像表現を用いた場合で 32.1%、q-GMM カーネルを用いた場合で 49.4%となった。ここで、q-GMM の q 値は q=1.05 を用いた。TRECVID 2010 データセットでは、q-GMM カーネルを用いた場合、Mean AP が 7.11%となり、q-GMM が通常の GMM(q=1) よりも高い精度を示した。

今後は、検出対象ごとの q 値の最適化や、q-GMM をさらに他の分野に応用することが課題として挙げられる。

5. 主な発表論文等
(研究代表者、研究分担者及び連携研究者には下線)

〔雑誌論文〕(計 2 件)

① Nakamasa Inoue, Koichi Shinoda, ″q-Gaussian Mixture Models for Image and Video Semantic Indexing″, Journal of Visual Communication and Image Representation, vol. 24, no. 8, pp. 1450-1457, Nov. 15, 2013 (査読有).
DOI: 10.1016/j.jvcir.2013.10.005

② Hilman F. Pardede, Koji Iwano, Koichi Shinoda, ″Feature normalization based on non-extensive statistics for speech recognition″, Speech Communication, vol. 55, pp. 587-599, Mar., 2013(査読有).
doi:10.1016/j.specom.2013.02.004

〔学会発表〕(計 5 件)

① Nakamasa Inoue, Zhuolin Liang, Mengxi Lin, Tran Hai Dang, Koichi Shinoda, Zhang Xuefeng, Kazuya Ueki, ″TokyoTech-Waseda at TRECVID 2014″, Proc. TRECVID workshop, pp. 1-13, Nov. 9, 2014 (米国、オーランド).

② Koichi Shinoda, ″(招待講演) Robust Video Information Retrieval using Speech Technologies″, Language Technologies Institute, Carnegie Mellon University, Jun. 20, 2014(米国、ピッツバーグ).

③ 周澤西, 岩野公司, 篠田浩一, ″音声認識のための q ガウス分布を用いた音響モデル″, 日本音響学会春季研究発表会, pp. 175-178, Mar. 13, 2013 (東京工科大学, 東京都八王子市).

④ Nakamasa Inoue, Koichi Shinoda, ″q-Gaussian Mixture Models Based on Non-Extensive Statistics for Image And Video Semantic Indexing″, ACCV2012, Nov. 5, 2012 (韓国, Daejeon).

⑤ Hilman F. Pardede, Koichi Shinoda, Koji Iwano, ″Q-Gaussian based spectral subtraction for robust speech recognition″, InterSpeech2012, Sep. 11, 2012(米国、ポートランド).

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況(計 0 件)

○取得状況(計 0 件)

〔その他〕
ホームページ等
http://www.ks.cs.titech.ac.jp/index.html

招待講演
Koichi Shinoda, ″Robust Video Information Retrieval using Speech Technologies″, Language Technologies Institute, Carnegie Mellon University, Jun. 20, 2014.

6. 研究組織
(1)研究代表者

篠田 浩一(SHINODA, Koichi)
東京工業大学・大学院情報理工学研究科・教授
研究者番号:10343097