様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25 年　6 月　18 日現在

機関番号：34406
研究種目：挑戦的萌芽研究
研究期間：2011 ～ 2012
課題番号：２３６５００７２
研究課題名（和文）　ML-BEATS 法の新たな展開

研究課題名（英文）　Searching new applications for ML-BEATS

研究代表者
鈴木 基之（SUZUKI MOTOYUKI）
大阪工業大学・情報科学部・准教授
研究者番号：３０２８２０１５

研究成果の概要（和文）：本研究では，音声符号化用に開発されたセグメント量子化法である ML-BEATS 法を他分野に適用し，その有効性を探った。まずは ML-BEATS 法を一般の時系列解析法として定義し直し，時間的遷移を考慮した話者識別用モデルの構築，音声認識・合成に基づく超低ビットレート音声符号化の 2 分野に適用した。ML-BEATS 法の実装に時間がかかったことから音声符号化については十分な検討が行えなかったが，話者識別モデルは従来の方法と比較して少ない学習データに対してよい性能を示すことがわかった。

研究成果の概要（英文）：In this study, we have applied the ML-BEATS method to two research fields, constructing a new speaker model and a new speech coding system based on speech recognition and synthesis. Experimental results showed that the new speaker model gave higher performance than conventional models when a few training samples can be used. On the other hand, we could not evaluate sufficiently a performance of a new speech coding system because of delay of implementation of ML-BEATS method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,200,000	660,000	2,860,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究開始当初の背景

ML-BEATS 法は，応募者が若手研究（B）（研究期間：平成 22 年度まで）において開発したセグメント量子化法である。この研究では，入力音声を ML-BEATS 法を用いてセグメント量子化することで，高効率な音声符号化法を実現した。ここで開発した ML-BEATS 法は音声信号以外にも，任意の時系列の入力が可能であり，この場合は時系列中からなんらかの「かたまり」を自動的に発見する方法，とみなすことができる。
一方，音声信号に限らず，脳波や株価といったように時系列信号は世の中に数多くあり，それらに対する解析法も様々に提案されているが，そうした方法の性能はいまだ十分ではなく，更なる解析法が望まれている。
そこで本研究では，ML-BEATS 法を時系列解析法として各研究分野に適用し，それぞれの分野における有効性を評価する。

２．研究の目的

本研究の目的は，音声符号化用に開発されたセグメント量子化法である ML-BEATS 法を他分野に適用し，その有効性を探ることである。
まずは ML-BEATS 法を一般の時系列解析

法として定義しなおし，具体的なツールキットも含めて開発を行う。その後，音声を対象とした以下の分野に適用し，それぞれの分野における適用可能性を探る。

- 時間的遷移を考慮した話者識別用話者モデルの構築
- 音声認識・音声合成に基づく超低ビットレート音声符号化

更に，得られた知見をもとに ML-BEATS 法の時系列解析法としての特徴を検討し，他分野へ適用する際の指針を得る。

３．研究の方法

若手研究（B）で開発した ML-BEATS 法は，HMM の学習とサンプルのパスへの割り当てを繰り返し行うため，大規模なモデルを学習しようとすると非常に時間がかかってしまう，という問題点がある。そこで，まず ML-BEATS 法のアルゴリズムを一部改良し，高速化を図る。また，プログラムを一部修正し，任意の時系列データを入力できるよう，ファイルフォーマット変換等の整備を行う。こうした整備を行った後，個別の問題に対してそれぞれ適用し，その性能を評価する。
具体的には，以下のように研究を行う。

（1）ML-BEATS 法の高速化とプログラムの整備

ML-BEATS 法を用いて音声符号化を行った研究の経験から，ML-BEATS 法は非常に計算時間がかかる（10 時間程度の音声データに対して計算時間が数週間から 1 ヶ月程度，など）ことがわかっている。その主な原因は，HMM の状態分割を繰り返すたびに，パラメータの再推定や部分系列への分割を行っているためである。
そこで，「状態分割の前後で状態とベクトルの対応関係は変化しない」という仮定を置くことで，パラメータの再推定なしに状態分割を繰り返すアルゴリズムへと変更する。この仮定は HMM の学習において比較的よく用いられているものであり，すでにその妥当性が検証されている。そこでこの仮定を ML-BEATS 法に導入し，高速化を図る。
更に，部分系列への分割計算も高速化する。現在は汎用 HMM 学習パッケージである HTK を利用しているが，専用プログラムを開発し，無駄な計算をはぶくことで高速化を図る。
最後に，一般の時系列解析が可能となるよう，入出力フォーマットの変換プログラム等を準備し，後日公開できるようパッケージ化するとともに，ドキュメント類を充実させておく。

（2）時間的遷移を考慮した話者識別用話者モデルの構築

話者識別用の音声コーパスを用いて，各話者ごとに話者モデルを構築する。音声データ（を特徴量に変換したもの）を ML-BEATS 法に入力することで，多数の（各「かたまり」に対応した）HMM が出力される。これらをそのまま話者モデルとして利用する。
従来から用いられている GMM による話者モデル，また，音素 HMM を連結した話者モデルも同じ学習データから構築し，話者識別性能を用いて評価を行う。この際，各モデルごとに状態数や混合数を変化させ，それらと性能との関係について検討する。
なお，音声コーパスは世界的によく用いられている “NIST Speaker Recognition Evaluation” を購入して用いる。このコーパスは世界的な話者識別コンテストに用いられているものなので，そのコンテストに参加した研究者の結果と直接比較をすることが可能である。

（3）音声認識・音声合成に基づく超低ビットレート音声符号化

音声認識・音声合成に基づく音声符号化法では，入力された音声を一度音声認識し，その発話内容のみを復号側へと送信する。復号側では送信された発話内容を音声合成し，音声波形へと戻している。この時，音声合成法として音声認識に用いた HMM のパラメータから音声波形を再生する方法を用いているため，再生される音声は，HMM を学習した音声に類似した声質となる。
この時，送信される発話内容は HMM を選択するためのシンボルとしてしか意味はなく，結果として言語との対応がとれている必要はない。そこで，ML-BEATS 法を用いて言語と関係のない「かたまり」を HMM でモデル化し，その ID を送信することで，より自然な音声符号化法の構築を目指す。
ML-BEATS 法で得られた「かたまり」に対応する HMM を音素 HMM のかわりに用い，あとは従来法と全く同じ方法で音声符号化を行う。また，前年度構築した話者モデルを流用することで，個人性の保存についても検討を行う。
符号化の性能を評価するため，従来法による音声符号化，また一般によく用いられている MELP に基づく音声符号化も行い，10 名程度の被験者に聞かせることで主観評価を行う。

４．研究成果

（1）ML-BEATS 法の高速化とプログラムの整備

ML-BEATS 法の高速化については，ML-BEATS

法において時間がかかる部分のひとつである，HMnet の状態分割ステップにおいて，「状態分割の前後で状態とベクトルの対応関係は変化しない」という仮定を置いたアルゴリズムへと変更を行った。具体的には，状態分割アルゴリズムを SSS-free で採用されている方法から，ML-SSS で採用されている方法へと変更した。
こうして実装したプログラムを用いて試験的に ML-BEATS 法を実行してみたところ，当初予定ほど高速化されていないことがわかった。その原因は，各学習サンプルのパスへの割り当て計算に時間がかかっている，ということであった。
ML-BEATS 法においては，
1. 状態分割による HMnet の詳細化
2. 各学習サンプルの「かたまり」への分割

という 2 つのステップを繰り返しながら，徐々に HMnet を詳細化し，「かたまり」を自動で発見していく。この 2 つ目のステップにおいて，現在の HMnet 内に存在するすべてのパス（いくつかの状態を持つ left-to-right 型 HMM）の中で，最も尤度の高いパスへと各学習サンプル（の一部の系列）を割り当てる，ということが行われる。
この計算は，Viterbi alignment と呼ばれる方法で実行することができるが，パスの数が増えてくると，そのすべてのパスに対して尤度計算を行う必要があるため，その計算に非常に時間がかかってしまう。
当初は，HMM の汎用 tool kit である HTK に含まれている，Viterbi alignment 計算プログラムを用いて計算していた。しかし，HMnet の状態分割が進み，形状が複雑になってくると，存在するパスの数が数千～数万といった単位になり，その結果計算時間のほとんどをこの計算が占める，ということになってしまった。
これに対し，

- 計算を HTK の汎用ルーチンではなく，C 言語を用いて独自に開発する
- ひとつの状態を多数のパスが共有していることから，各状態での尤度を先に計算しておき，それをキャッシュとして各パスの尤度を計算する
- 各パスにおいて，途中までの尤度が低ければ計算を打ち切る，といったビームサーチの方法を導入する

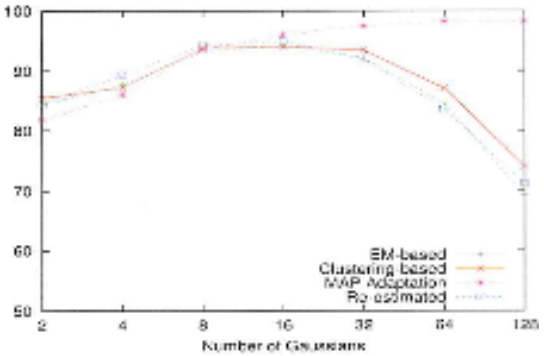
といった対策をした，高速版プログラムの開発を試みた。
しかし，このプログラムの開発が思いの他難航し，結果として研究期間内に完成させることができなかった。その主な原因は，HTK フォーマットの HMM 定義ファイルの扱いに問題があったこと，またビームサーチの導入に関し，アルゴリズムの確定や実装方法の検討に時間がかかってしまったことである。
しかし，おおよその目途は立ちつつあることから，本研究課題終了後も，本補助金で購入した計算機を活用しながら開発を続け，近い将来には toolkit として公開できるよう努力していく予定である。

（2）時間的遷移を考慮した話者識別用話者モデルの構築
本テーマについては，ML-BEATS 法のプログラムが完成しなかったことから，当初予定していた時間的遷移を考慮した話者モデルの構築を行うことはできなかった。そこで，ML-BEATS 法の考え方を取り込んだ，クラスタリングに基づく GMM 学習法による話者モデルの構築法を提案し，その性能を評価した。
現在の話者識別において，話者モデルは GMM（Gaussian mixture model）がよく用いられている。これは 1 つの確率分布を多数の正規分布の重み付き和で表現しよう，というモデルであり，様々な発話内容が含まれる音声データを表現するのによいモデルとなっている。しかし，正規分布の数を増やしていくと，その一部は少ない学習サンプルに非常に特化したものとなったり，また別の一部は全体をカバーする非常に汎用的なのになったりと，学習がうまく進まずによいモデルとはならなくなる。
もともと多数の正規分布が必要であった理由は，多様の発話をカバーするためである。そのため，個々の正規分布は多様な発話の一部にそれぞれ対応していることが自然であると思われる。ML-BEATS 法では，複数のパスに学習サンプルを分割することで，一部の学習サンプルを一部のパスに，と明示的に対応関係を求めながら学習されていく。その結果よいモデルが得られていた。そこで，各正規分布を特定の発話と明示的に対応づけ，その特徴が消されないようにするアルゴリズムを提案した。
具体的には，すべての学習サンプルを用いて k-means 法でクラスタリングを行い，それぞれをひとつの正規分布に対応させる。k-means 法を行う際には，そのセントロイドはベクトルではなく，所属している全学習サンプルから計算した正規分布を用い，また距離計算も正規分布から計算される尤度を用いた。こうすることで，正規分布の尤度（認識時の基準と同じ）を用いて適切なクラスタリングが行われることとなる。その結果，特定の特徴を消さない正規分布が学習され，よりよい話者モデルが得られることが期待される。
提案した方法の性能を評価するため，話者識別実験を行った。登録話者は男女 50 名ずつの 100 名，登録には話者 1 名あたり 2 文を用いた。評価に用いた文は各話者あたり 20 文である。また，GMM の混合数は 2～128 とした。

性能を比較するため，従来法として，

- 最尤推定した GMM
- 不特定話者 GMM を MAP 適応で作成した GMM
- クラスタリングによる GMM（提案方法）のパラメータを最尤推定し直した GMM

についても同様の実験を行った。
2 文だけを用いて話者を登録した時の結果を図に示す



これを見ると，混合数が少ない時（16 混合程度まで）は，どの方法もほとんど差がないことがわかる。一方，32 混合以上になると，MAP 推定を用いた方法は性能を向上させているが，その他の方法は性能を劣化させていることがわかる。これは，MAP 推定の方法だけが，登録話者とは別の 100 名を用いて不特定話者 GMM を学習させていることから，正規分布の数が増えても過学習を起こさずにいたことが原因と思われる。
性能が下がってしまった方法のうち，提案方法はその低下割合が一番低く，3 つの方法の中では最もよい性能を示していた。これはクラスタリングに基づいていることで，個々の正規分布が学習サンプルと明示的に対応づけられ，極端に特化した，といったことにならなかったため，過学習の影響が少なくなったためと思われる。
今回は ML-BEATS 法を用いることができなかったため，性能向上の効果は限定的なものであった。しかし，ML-BEATS 法の考え方が有効であることは示せたため，今後時間的遷移も考慮した（ML-BEATS 法による）話者モデルを構築できれば，より高い性能となることが期待される。

（3）音声認識・音声合成に基づく超低ビットレート音声符号化

現在携帯電話等で用いられている音声符号化法は，音声を LSP 係数等に変換し，その値を送信することで低ビットレート化している。これに対し，一度音声認識を行い，その結果を送信して，復号側では音声合成を用いることで，より低ビットレートな音声符号化が可能であることが示された。
この方法では，「音素」が符号として用いられている。しかし，音声合成も HMM に基づく方法を用いているので，「音素」を符号とする必要はなく，「HMM で表現されているもの」であれば，同じ枠組みで符号化することができる。
この時，「HMM で表現されているもの」とは，音声の特徴をよく表現したものであり，その性質はなるべく似ているものが集まって構築された HMM であることが望ましい。そこで，ML-BEATS 法を用いて音声をクラスタリングし，HMM で表現することで，より低ビットレートな音声符号化の実現を目指した。
音声符号化，復号化のプログラムや，比較対象としての音素を用いた音声符号化法については，準備を行っていたが，ML-BEATS 法を用いて大規模な HMM モデルを作成することができなかったため，実際に音声符号化実験を行えずに研究期間が終了してしまった。実験についてはその準備は整えているため，研究期間終了後も引き続き検討を行う予定である。

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 3 件）

1. Motoyuki Suzuki, Shohei Nakagawa, and Kenji Kita. “Prosodic feature normalization for emotion recognition by using synthesized speech,” Proc. 16th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems, pp.306–313. (2012) （査読有）
2. Motoyuki Suzuki, Seiji Tsuchiya, and Fuji Ren. “A novel emotion recognizer from speech using both prosodic and linguistic features,” Proc. 15th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Vol. I, pp.456–465. (2011) （査読有）
3. 御船 正樹, 鈴木 基之, 任 福継, 北 研二. 「クラスタリングに基づく GMM 学習法による話者のモデル構築」, 電子情報通信学会 音声研究会資料 SP2011-42 (2011) （査読無）

〔学会発表〕（計 3 件）

1. Motoyuki Suzuki, “Prosodic feature normalization for emotion recognition by using synthesized speech,” 16th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2012 年 9 月 10 日, サンセバスチャン(ス

ペイン）

2. Motoyuki Suzuki, “A novel emotion recognizer from speech using both prosodic and linguistic features,” 15th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2011 年 9 月 13 日, カイザースラウテルン（ドイツ）
3. 御船正樹,「クラスタリングに基づく GMM 学習法による話者のモデル構築」電子情報通信学会 音声研究会, 2011 年 7 月 21 日, 札幌市

６．研究組織

(1)研究代表者

鈴木 基之（SUZUKI MOTOYUKI）

大阪工業大学・情報科学部・准教授

研究者番号：30282015

(2)研究分担者

なし

(3)連携研究者

なし

(4)研究協力者

御船 正樹（MIFUNE MASAKI）

徳島大学・大学院先端技術科学教育部・博士前期課程 2 年