論文 / 著書情報
Article / Book Information

論題(和文)	VQひずみ、離散/連続HMMによるテキスト独立形話者認識法の比較検討
Title(English)	
著者(和文)	松井知子, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会論文誌, Vol. J77-A, No. 4, pp. 601-606
Citation(English)	, Vol. J77-A, No. 4, pp. 601-606
発行日 / Pub. date	1994, 4
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1994 Institute of Electronics, Information and Communication Engineers.

論　文

VQひずみ，離散/連続HMMによるテキスト独立形話者認識法の比較検討

正　員　松井　知子[†]　　正　員　古井　貞熙[†]

Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs

Tomoko MATSUI[†] *and* Sadaoki FURUI[†], *Members*

あらまし　VQ（ベクトル量子化）ひずみ，および離散/連続エルゴード的HMM（隠れマルコフモデル）による話者認識実験の比較を行い，発声変動（発声内容の違い，時期的な変動，発声速度の変化）に対する頑健性の観点から，それらの手法の有効性について検討する．連続HMMの性能は離散HMMに比べてはるかに良いこと，また，VQひずみの性能とほぼ等しいことを示す．特に，学習データ量が十分でない場合には，VQひずみの方が連続HMMよりも認識率が良い．状態間の遷移情報は個人差の表現にほとんど寄与しないため，状態数の増加と状態内の混合分布数の増加は同じ効果があり，認識性能はほぼ両者の積によって決定されることを示す．

キーワード　話者認識，テキスト独立，ベクトル量子化，エルゴード的HMM，発声変動

1. まえがき

VQ（ベクトル量子化；vector quantization）ひずみによる方法は，テキスト独立形話者認識において，標準的な手法として利用されている．一方，HMM（隠れマルコフモデル；hidden Markov model）も，ここ数年よく使われるようになってきた．Rosenberg[1]は，left-to-right HMMによる話者認識について，Poritz[2]とTishby[3]は，linear predictive ergodic HMMによる話者認識について報告している．また，Savic[4]は，エルゴード的HMMの各状態に割り振られるテストサンプルと学習サンプルの類似度によって話者照合を行った．更に，松本[5]は，テキスト独立形話者認識においてスペクトル空間を大分類すれば，音韻情報に依存した話者情報のばらつきを軽減できることを示した．エルゴード的HMMは，各状態に対応して，スペクトル空間の大分類が自動的になされるので，テキスト独立形話者認識に有効であるとされている．しかし，エルゴード的HMMのゆう度を，直接話者の判定に利用した研究例の報告はなく，テキスト独立形話者認識における離散，連続HMMの性能の違いも明らかでない．

本論文では，VQひずみ，離散/連続エルゴード的HMMによる話者認識実験を通して，それらの手法の比較を行い，発声変動（発声内容の違い，時期的な変動，発声速度の変化）に対して頑健な話者認識法について検討する．

2. 話者認識法

VQひずみによる話者認識では，話者ごとに符号帳を用意し，入力音声データと各々の符号帳との平均量子化ひずみに基づいて話者を判定する[6]．

HMMによる話者認識では，話者ごとにエルゴード的HMMを用意し，入力音声データを各々のエルゴード的HMMに与えたときのゆう度に基づいて話者を判定する．図1は，4状態のエルゴード的HMMの例である．ここでは，離散HMMとしてFuzzy-VQ HMMを，連続HMMとして混合無相関ガウス分布

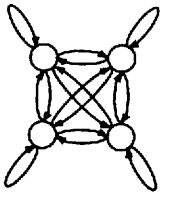

図1　状態のエルゴード的HMM
Fig. 1　4-state ergodic HMM.

† NTTヒューマンインタフェース研究所，武蔵野市
NTT Human Interface Laboratories, Musashino-shi, 180 Japan

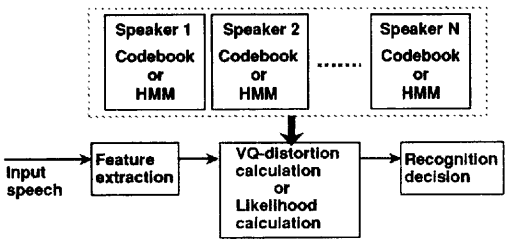

図 2 話者認識システムの構成
Fig. 2 Speaker recognition system block diagram.

HMM を使用した(図 2).

3. 実　　験

3.1 概　　要

使用したデータは，男 23 名，女 13 名が約 5 か月にわたる 3 時期に，3 種類の速度（普通，速い，遅い）で発声した文章データである．文章テキストは，ATR 連続音声テキスト 503 文[7] から抜粋した．各話者は，そのテキストを読み上げる形で発声した．なお，録音は雑音のない部屋で行われ，全話者，全時期で同じコンデンサーマイクロホンを使用した．

特徴パラメータとして，標本化周波数 12 kHz，フレーム長 32 ms，フレーム周期 8 ms，LPC 分析次数 16 でケプストラムを抽出した．その際，音声パワーは用いなかった．表 1 に示すように，学習にはある 1 時期に普通速度で発声した 10 文章を用いた．テストでは，その他の 2 時期に普通，速い，遅い速度で発声した 5 文章を 1 文章ずつ用いた．学習，テストデータの発声時期の組合せは，3 時期について順ぐりに交替させた．学習で用いた 10 文章のうちの 5 文章は，そのテキストが全話者，全時期に共通で，残りの 5 文章のテキストは各話者，各時期ごとに異なる．テストで用いた文章は，学習に用いた文章とは異なるが，そのテキストは全話者，全時期で同じである．普通，速い，遅い速度で発声した文章の平均時間長は，それぞれ 4.2 秒，3.2 秒，5.8 秒であった．実験に用いたテスト音声の総数は，3 種類の各発声速度に対して，36(名)×2(時期)×3(組合せ)×5(文章)＝1,080(文章)である．

VQ ひずみ，HMM による話者認識法の性能は，話者識別率，話者照合率で評価した．話者識別の場合は，36 名の話者の中から最も類似度の大きい話者（VQ ひずみの場合は平均量子化ひずみの最も小さい話者，HMM の場合はゆう度の最も大きい話者）を選択した．話者照合の場合は，類似度の値にしきい値を設けて，しきい値よりも類似度の大きいテストデータは本人の音声として受理し，それ以外は棄却した．このしきい値は，話者ごとに 2 種類の誤り（詐称者受理率，本人棄却率）が等しくなる値に事後的に設定した．なお，しきい値は各話者ごとに設定した．

表 1 音声データ

		文章テキスト
学習	1.	飛ぶ自由を得ることは人類の夢だった。
	2.	初めてルーブル美術館へ入ったのは１４年前のことだ。
	3.	自分の実力は自分が一番よく知っているはずだ。
	4.	これまで少年野球ママさんバレーなど地域スポーツを支え市民に密着してきたのは無数のボランティアだった。
	5.	ギンザケの卵を輸入してふ化させ海中で育てる養殖も始まっている。
	(6).	()の５つのテキストについては、各話者、各時期毎に異なるものを、ATR 連続音声テキスト５０３文から抜粋して使用した。
	(7).	
	(8).	
	(9).	
	(10).	
テスト	1.	予防や健康管理リハビリテーションのための制度を充実していく必要があろう。
	2.	背の高さは１７０センチ性どで目が大きくやや太っている。
	3.	大声を出し過ぎてかすれ声になってしまう。
	4.	足し算引き算はできなくても絵は描ける。
	5.	どの部屋の意匠にも遊び心があふれていて楽しい。

VQ ひずみによる実験では，符号帳サイズを 32 から 512 まで変化させた．離散 HMM による実験では，符号帳サイズを 256 から 1,024 まで変化させ，また状態数は符号帳サイズが 256 の場合には 1 から 8 まで変化させ，1,024 の場合には 4 とした．連続 HMM による実験では，混合ガウス分布数を 8 から 128 まで変化させ，状態数についても 1 から 8 まで変化させた．

VQ 符号帳の作成には，LBG アルゴリズムを用いた．HMM の出力シンボル確率の初期化は，離散 HMM では学習データを状態数で分割し，各々についてのコードワードの頻度分布により，連続 HMM では学習データを総分布数（状態数 × 混合分布数）で分割し，各々についての平均ベクトル，共分散を求めて行った．遷移確率の初期化については，各状態から出る遷移を等確率とした．また，HMM パラメータ推定において，各状態から出る遷移の出力シンボル確率は同一とした．HMM のパラメータは Baum-Welch アルゴリズムで推定した．

3.2 結　　果

3.2.1 話者識別

識別結果を図 3 に示す．この実験では，1 状態 64 混合分布の連続 HMM が最も性能が良かった．特に速い速度で発声されたデータに対して良かった．全般的に，連続 HMM による方法の性能は VQ ひずみによる方法とほぼ等しく，また離散 HMM による方法は非常に性能が悪い．

VQ ひずみによる方法では，符号帳サイズが大きくなるにつれて，識別率も上昇した．連続 HMM による方法では，1 状態 128 混合分布の連続 HMM は 1 状態 64 混合分布よりも性能が悪い．これは，混合するガウス分布数が多くなると，各分布の分散を推定することが難しくなるためと考えられる．離散 HMM による方法に関しては，状態数を変化させてもほとんど識別率は変わらなかった．また，符号帳サイズが 256 の場合と 1,024 の場合との比較からわかるように，符号帳サイズを更に大きくすれば，離散 HMM の性能も向上する可能性もあるが，計算量および学習データ量が膨大になる問題がある．

発声速度に関しては，普通速度で発声されたデータが最も識別率が高く，速い速度で発声されたデータが最も低かった．学習には普通速度の音声を用いたことを考慮すると，このことは速い速度で発声されたテストデータの分布が，普通と遅い速度で発声されたデータの分布から離れていることを示している．

3.2.2 話者照合

照合結果を図 4 に示す．この実験では，符号帳サイズ 512 の VQ ひずみが最も性能が良かった．しかしながら，全般的に，連続 HMM と VQ ひずみとの性能の

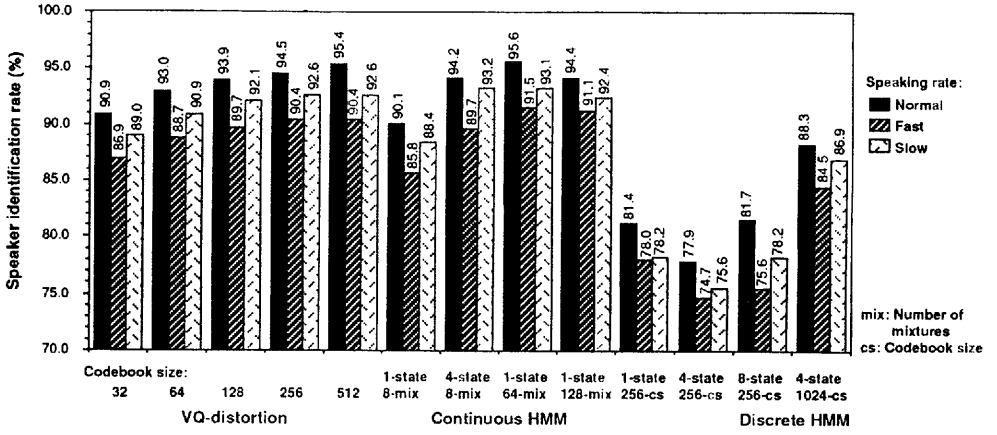

図 3　話者識別率(%)
Fig. 3　Speaker identification rate (%).

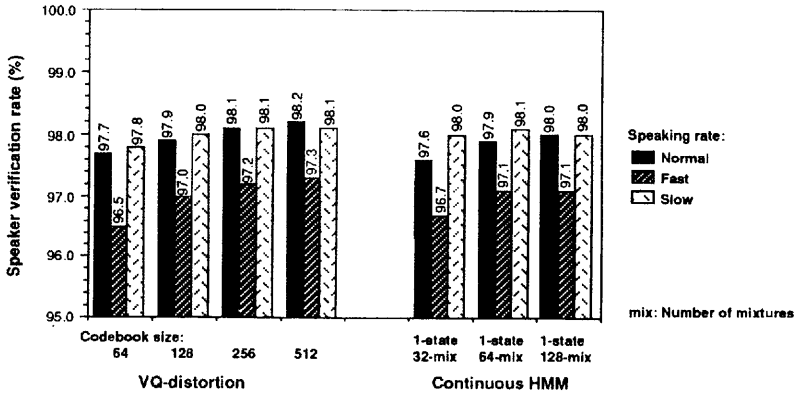

図 4　話者照合率(%)
Fig. 4　Speaker verification rate (%).

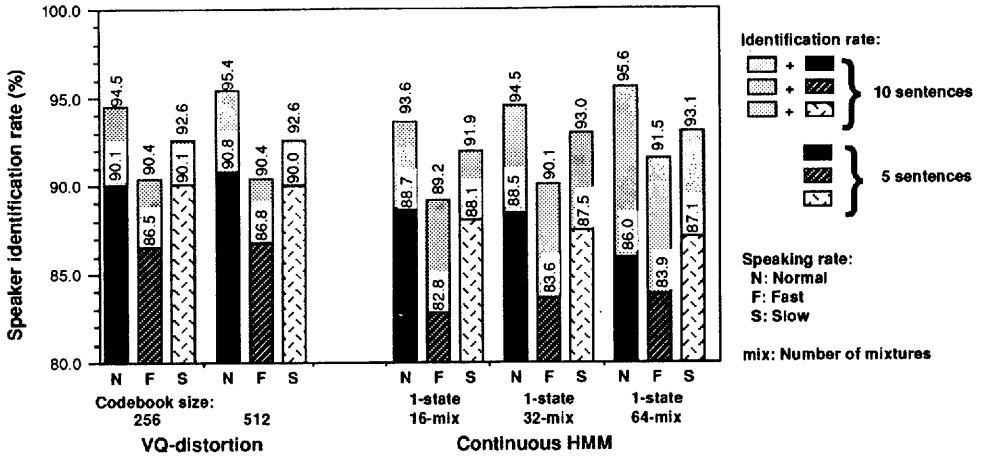

図 5 異なる学習データ量による話者識別率
Fig. 5 Speaker identification rates with different amounts of training data.

差はほとんどない.

発声速度に関しては，話者識別の場合と同様に，速い速度で発声されたデータは，普通と遅い速度で発声されたデータに比べて照合率が低い.

3.2.3 学習データ量との関係

VQ ひずみ，連続 HMM による方法において，学習データ量との関係を調べるために，学習データ量を半分（5 文章）にして，話者識別実験を行った．ここでは，表 1(6)〜(10)の各話者，各時期ごとに異なる 5 文章を学習に用いた．VQ ひずみによる方法では，サイズが 256，512 の符号帳を，連続 HMM による方法では，1 状態，16，32，64 混合分布の HMM を用いた．結果を，学習に 10 文章を用いた場合と併せて，図 5 に示す．図 5 より，VQ ひずみの方が連続 HMM よりも，学習データ量の減少に対する認識率の低下が少ないことがわかる．また，連続 HMM に関して，混合分布数 32 と 64 の場合を比べると，学習に 10 文章を用いた場合は，32 から 64 へ混合分布数を増加させることによって性能の向上が見られたが，この実験結果では普通速度と遅い速度の発声に関して，識別率にむしろ低下が見られる．連続 HMM では，学習データ量が少ない場合には，HMM パラメータの推定がうまくできず，VQ ひずみのようなノンパラメトリックな方法に比べて，影響を受けやすいと考えられる．少ない学習データ量で連続 HMM のパラメータを頑健に推定するためには，分散の値に制限を設ける等の制約が必要であろう.

VQ ひずみに関しては，学習データとして 10 文章を用いた場合，符号帳サイズ 512 が識別率が最も高かったが，学習データとして 5 文章を用いた場合には，識別率は符号帳サイズ 256 でほぼ飽和した．それぞれの飽和レベルにおける，1 クラスタ当りの平均学習ベクトル数はともに 11 であった．前述のように連続 HMM に関しては，学習データとして 10 文章を用いた場合には，1 状態 64 混合分布の HMM が最も識別率が高かった．学習データとして 5 文章を用いた場合には，識別率は 1 状態 32 混合分布のときにほぼピークを示した．それぞれのピークレベルにおける，1 ガウス分布当りの平均学習ベクトル数はともに 89 であった．これらの値は，VQ ひずみあるいは連続 HMM を用いる場合に必要となる学習サンプル量の目安を与えていると考えられる.

4. 考　　察

4.1 離散/連続 HMM の性能の違い

離散，連続エルゴード的 HMM の性能の違いについては，次の二つの要因が考えられる.

[要因 1] 図 6 に示すように，離散エルゴード的 HMM では，あるテストベクトルの出力シンボル確率は，それに一番近い VQ コードワードの出力シンボル確率に設定される．テキスト独立形話者認識においては，使用可能なデータが少なく，しかも発声変動を含むデータを扱った場合，テストデータと学習データの分布のずれが大きくなる．あるテストデータを他人の HMM に与えて，そのゆう度を計算するときに，学習データの分布から外れたテストベクトルに対応する VQ コードワードの出力シンボル確率がたまたま高い

と，そのテストデータは誤認識され，識別率は低下する．連続 HMM においては，そのようなテストベクトルは，ガウス分布の端に相当するので，その出力シンボル確率は低くなる．

これは，連続 HMM によるゆう度がガウス分布で表され，テストベクトルが分布から離れるにつれて滑らかに減少する関数であるのに対し，離散 HMM によるゆう度が頻度分布で表され，テストベクトルと分布との距離に関係なく，階段状の値をもつ関数であることに起因する．このことが連続エルゴード的 HMM と離散エルゴード的 HMM の性能の違いに対応していると考えられる．

なお，VQ ひずみを直接用いる方法の場合は，コードワードとの距離値が滑らかな評価関数として用いられるので，1 状態の連続 HMM の場合に近い結果を与えると考えられる．

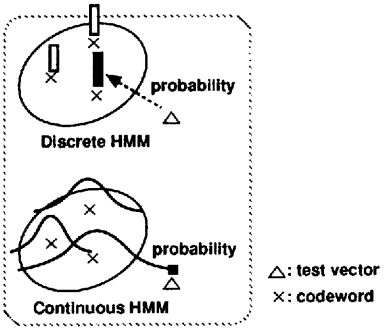

図 6　離散，連続 HMM における出力シンボル確率の計算方法
Fig. 6　Illustration of discrete HMM vs. continuous HMM.

[要因 2]　離散 HMM において，符号帳サイズを大きくすると性能が向上することから，量子化誤差に話者情報が埋もれていて，話者の判定性能が低くなっていることが考えられる．従って，符号帳サイズを更に大きくすれば，性能も向上する可能性もあるが，3.2.1 で述べたように計算量および学習データ量が膨大になる問題が生じ，パラメータ推定が難しくなる．

4.2　連続 HMM によるスペクトル空間の大分類の効果

連続エルゴード的 HMM について，その状態数と，混合分布を構成する分布数を変えて話者識別実験を行い，スペクトル空間を大分類する効果を調べた．図 7 に示すように，すべての発声速度に対して，状態数，分布数が増えれば，識別率も上がるが，識別率はほぼ総分布数（状態数 × 分布数）によって決定されることがわかる．速い発声速度の場合を除くと，総分布数に対する識別率は，総分布数 32 でほぼ飽和する傾向にある．これらの結果は，次の二つの可能性を示唆している．

第 1 に，異なる状態間の遷移情報は，個人差の表現にほとんど寄与しない可能性がある．状態間の遷移，すなわち音素クラス間の遷移は，主にテキストに依存するものであり，特にテキスト独立な条件のもとでは，個人差を表す情報としての有効性は少ないので，状態数の増加と状態内の混合する分布数の増加は同じ効果になると考えられる．

第 2 に，遷移確率に個人差があったとしても，その

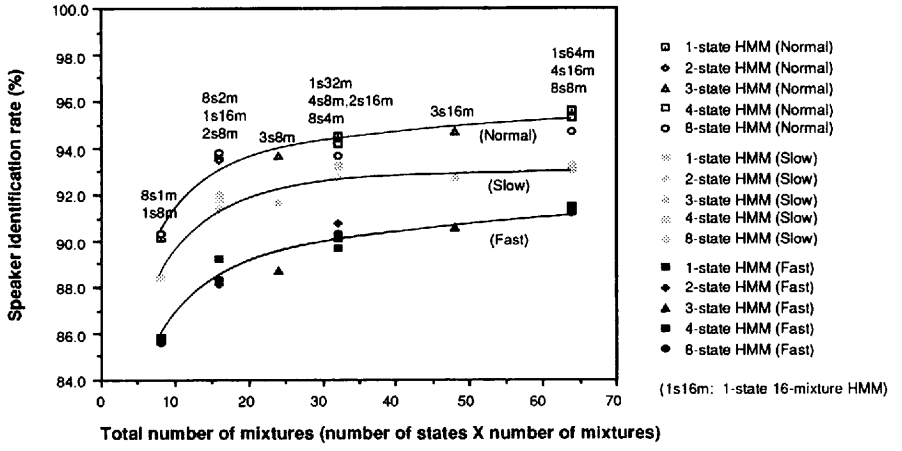

図 7　状態数，混合分布数を変化させたときの話者識別率
Fig. 7　Speaker identification rates as functions of the numbers of states and mixtures.

値が非常に小さいので（本実験においては，HMM学習後の異なる状態間の遷移確率は，すべて0.1から0.2），ゆう度の計算に影響を与えない可能性がある[8]．

5. むすび

テキスト独立形話者認識において，VQひずみ，離散/連続エルゴード的HMMによる方法を比較し，連続HMMは離散HMMに比べてはるかに性能が良いこと，また，VQひずみと比べるとほぼ等しいことを示した．特に，学習データ量が十分でない場合には，VQひずみの方が連続HMMよりも認識率が良い．状態間の遷移情報は個人差の表現にほとんど寄与しないため，状態数の増加と状態内の混合分布数の増加は同じ効果があり，認識性能はほぼ両者の積によって決定されることを示した．

今後は，音韻性情報の有効な利用方法などについて検討していく予定である．そのときには，音声認識との整合性から，VQひずみよりもHMMを用いた方が有効であると考えられる．更に動的特徴を用いた実験も検討していきたい．

謝辞 熱心に討論して頂いた，ヒューマンインタフェース研究所古井特別研究室の方々に感謝致します．

文　献

(1) Rosenberg A. E., Lee C.-H. and Cokcen S.: "Connected word talker verification using whole word hidden Markov models", Proc. ICASSP, pp. 381-384 (1991).
(2) Poritz A. B.: "Linear predictive hidden Markov models and the speech signal", Proc. ICASSP, pp. 1291-1294 (1982).
(3) Tishby N. Z.: "On the application of mixture AR hidden Markov models to text independent speaker recognition", IEEE Trans. ASSP, pp. 563-570 (1991).
(4) Savic M. and Gupta S. K.: "Variable parameter speaker verification system based on hidden Markov modeling", Proc. ICASSP, pp. 281-284 (1990).
(5) Matsumoto H.: "Text-independent speaker identification from short utterance based on piecewise discriminant analysis", Comput. Speech Lang., pp. 133-150 (1989).
(6) Soong F. K., Rosenberg A. E., Rabiner L. R. and Juang B. H.: "A vector quantization approach to speaker recognition", Proc. ICASSP, pp. 387-390 (1985).
(7) 桑原尚夫，匂坂芳典，武田一哉，阿部匡伸："研究用ATR日本語音声データベースの作成（別冊I連続音声テキスト）", TR-I-0086 (1989).
(8) 清野　崇，中川聖一："エルゴディックHMMを用いた音声による多言語間の識別", 信学技報, **SP92**-129 (1993).

(平成5年9月2日受付，11月24日再受付)

松井　知子

昭61東工大・理・情報卒．昭63同大大学院修士課程了．同年NTT(株)入社．以後，NTTヒューマンインタフェース研究所において，話者認識の基礎研究に従事．IEEE，日本音響学会各会員．平5論文賞受賞．

古井　貞熙

昭43東大・工・計数卒．昭45同大大学院修士修士課程了．同年NTT電気通信研究所入社．以後，同研究所において，音声認識，話者認識，音声知覚などの研究に従事．昭53～54ベル研究所客員研究員．現在NTTヒューマンインタフェース研究所古井特別研究室長．工博．昭50米沢賞，昭63，平5論文賞，平2著述賞，昭60，62音響学会論文賞，平1科学技術庁長官賞，IEEE ASSP Society Senior Award各受賞．著書「ディジタル音声処理」，「Digital Speech Processing, Synthesis, and Recognition」，「音響・音声工学」，「音声情報工学」(分担)，「音の科学」(分担)，編著「Advances in Speech Signal Processing」など．IEEE Fellow．ESCA，日本音響学会各会員．