口ドラム認識手法とそのドラム譜入力システムへの応用

中野倫靖[†] 緒方淳[††]
後藤真孝[††] 平賀譲[†]

本論文では，人がドラムの音を「ドンタンドドタン」のように擬音語で真似た音声（口ドラム）を，それに対応するドラムパターンとして認識する手法を提案する．口ドラムには，声質などの発声スタイルの個人差と表現の個人差の 2 種類の個人差が存在し，認識手法はこれらの個人差を吸収できるものが望ましい．従来，実際のドラム音（楽器音）を対象とした認識は研究されてきたが，それらの手法では口ドラムの多様な個人差への対処が難しかった．そこで本手法では，口ドラムを，その各音を表す音素列の擬音語で表現し，多様な擬音語表現の辞書を用意することで，表現の個人差に対処する．さらに，音声認識で用いられている音響モデルを用いることにより，各歌唱者間の声質の個人差を吸収する．本手法を 200 発話の口ドラムデータに対して適用した結果，最も良い実験条件において 92.0%の認識率を得た．この結果は，提案手法が十分実用性を持つことを示しており，応用例として，口ドラムによるドラム譜入力インタフェース Voice Drummer を実装した．

A Voice Percussion Recognition Method and Its Application to a Music Notation System of Drum Sounds

Tomoyasu Nakano,[†] Jun Ogata,[††] Masataka Goto[††]
and Yuzuru Hiraga[†]

This paper presents a method of recognizing voice percussion (verbalized expression of drum sound by voice) as an expression of intended drum patterns. Recognition of voice percussion requires an approach that is different from existing methods for drum sound recognition. Individual differences in both vocal characteristics and the verbal expressions used add further complication to the task. The approach taken in this study uses phonemic sequences of onomatopoeia as internal representation of drum sounds. The set of onomatopoeia used in drum sounds are included in a pronunciation dictionary, and the phonemic sequences are estimated by utilizing an acoustic model. The acoustic model and the dictionary are intended to deal with the two types of individual differences mentioned above. In a recognition experiment with 200 utterances of voice percussion, our method achieved a recognition rate of 92.0% for the highest-tuned setting. Following the results of the proposed method, *Voice Drummer*, a music notation interface of drum sounds, was implemented, as a practical application for voice percussion recognition.

1. はじめに

本研究では，ドラム音を真似た「ドンタンドドタン」のような音声（口ドラム）を，ドラムパターンとして認識する手法を提案する．口ドラムとは，ドラム音を音声で表現することを指し，ヴォイスパーカッション（Voice Percussion）や，ビートボクシング（Beatboxing）などと呼ばれることもある☆．一般的には，これらは実際のドラム音と聞こえの差が小さい歌唱音声を指すことが多いが，本研究ではドラム音と音響的に似ていなくても，冒頭の擬音語表記のようにドラム音をイメージできるような音声であれば口ドラムと呼ぶ．また，ドラムパターンとは，1 小節を最小単位とするドラムスの演奏パターンを指す．

口ドラムの認識によってドラムパターンを得ることができれば，ドラムスの演奏経験がない場合でも直感的かつ手軽にドラムパターンを表現でき，これまで実

† 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
†† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

☆ そのほか，Human Beat Box，Vocal Percussion，Mouth Drumming など様々な名称がある．

現できなかった様々なアプリケーションの構築が可能となる．たとえば，作曲・編曲におけるドラムパートを既存ドラムパターン（ドラムループ）の素材集から選択する場面において有用である．また，楽曲検索の対応範囲を広げ，ドラムパターンが印象的な曲の検索にも利用しうる．たとえば，イントロが特徴的なドラムパターンで始まる楽曲や，同じドラムパターンが繰り返しているような楽曲の検索に適用できると考えられる．さらに，ギターなどの他の楽器と併用することで，One Man Band のような 1 人セッションを支援するアプリケーションが考えられる．

音声（歌唱）から対応メロディや楽曲を得る研究としては，ハミング（鼻歌）や歌声を用いた音楽検索手法がある[1)~3)]．これらは主として，検索キーに音高・音長情報を用いており，音色に関する情報は考慮していない．しかし，ドラムパートは，音色の違いこそが重要な要素であるため，これらの研究で得られた知見を，本研究にそのまま適用することは難しい．そこで，ドラム音を表現する手段として口ドラムに着目し，その認識を行う．

口ドラム認識に関連する研究として，ドラム音認識（音源同定）があり，これまで，ドラム単音を対象としたもの[4)]，ドラムパートの演奏を対象としたもの[5)]，ドラム音以外の複数の楽器音が混在した音響信号を対象としたもの[6),7)] がある．ドラム音認識に共通する問題はドラム音の個体差に関するものであり，その対処法としては，1 つのテンプレートを対象曲中で適応的に変化させる手法が提案されている[6),7)]．

近年増えてきた口ドラム認識の研究は，対象とする口ドラムの音響的性質によって 2 つに分けられる．1 つはドラム音を音響的に模倣した Beatboxing を認識する研究[8),9)]，もう 1 つは，本研究でも対象としている擬音語による発声を認識する研究[10),11)] である．前者が扱っている音は，楽器のドラム音と聞こえの差が小さいもの，すなわち訓練された Beatboxer などが発声した音声であり，訓練されていない歌唱者が発声することは困難である．また，純粋な音声だけでなく，マイクを手で覆った発声などのテクニックを用いた音も含まれる．そのため，本研究とは入力として扱う音声の性質が異なる．

後者のタイプの研究として，Gillet らは，ドラム音に対応する擬音語（Bass Drum としての boom や，Snare Drum としての tcha など）を決定し，それぞれの擬音語（口ドラム）の発声とそのスペクトル構造の対応に確率モデル（音響モデル）を用いることで認識を行っている[11)]．Gillet らの手法は，口ドラムを擬音語で表現してはいるが，それぞれを 1 つの音響モデルに対応させているため（たとえば，boom/tcha でそれぞれ 1 つずつ），採用されなかった擬音語☆で発声された口ドラム音を認識するためには，新たにモデルを構築する必要がある．それに対して本手法[10)] では，口ドラム音を擬音語として抽象化することを生かし，音響モデルとして音素単位のモデル（/d/や/o/など）を用いて認識を行う．ここで，音声認識で用いられる既存の（学習済み）モデルを採用することで，大量の学習データを前提とせずに多様な表現に対処できる．

以下，2 章で口ドラム認識手法の概要を述べる．3 章で口ドラムの擬音語表現の調査結果，4 章で提案手法の認識実験の結果を述べ，結果について考察する．そして 5 章では，本研究の提案手法の応用例であるドラム譜入力インタフェース Voice Drummer について説明し，最後に 6 章でまとめと今後の展望について述べる．

2. 口ドラム認識手法

本章では，口ドラム認識における課題である個人差について論じ，擬音語を用いた解決法を提案する．次いで，実際の認識手法について説明する．

2.1 口ドラム認識における課題

口ドラムを認識するためには，声質の個人差と表現の個人差という，2 種類の個人差への対処が必要になる．以下，それぞれの個人差の対処への課題を論じ，各ドラム音と口ドラム表現の中間形式として擬音語を採用することで，個人差に対処する方法を述べる．

(1) 声質の個人差

口ドラムには，性別の違いによる声の高低など，声質に個人差がある．口ドラムの声質の個人差への対処は，従来のドラム音認識における個体差の問題に相当する．それらの場合と同様に，すべての声質の口ドラムを収録したようなデータベースの構築によって対処するのは，現実問題として難しい．そのため，声質の違いを吸収するような認識手法が望ましい．

そこで，多数の話者の音素がどのようなスペクトル構造になるかを学習した確率モデルを用いることで，この問題を解決する．確率モデルとしては，不特定話者の音声認識に用いられる学習済みのモデル（音響モデル）を，口ドラム認識のスタイルに適応させて利用する．これにより，擬音語の各音素と口ドラム音のスペクトル構造とを対応付けることが可能となり，口ド

☆ Gillet らが採用した擬音語は Bass Drum で 2 つ（boom, poom），Snare Drum で 1 つ（tcha）であった[11)]．

ラム歌唱における声質の差異を吸収できる．

本論文では，擬音語音素の音響モデルとして，連続音声認識コンソーシアム（Continuous Speech Recognition Consortium: CSRC）で頒布されている，2002年度版の不特定話者 HMM（Hidden Markov Model）を使用した[12]．これは，男女約 270 名，約 40,000 発話から学習された性別非依存（GID）モデルであり，モデルのタイプは monophone HMM（音素コンテキスト独立型）である．

(2) 表現の個人差

ドラム音を耳で聴いたときのイメージは歌唱者ごとに異なるため，それを歌った口ドラムも，「ドンタン」や「ズンチャ」のように，異なった表現が使用されると考えられる．さらに，同一の歌唱者であっても，場合により異なる表現を用いる可能性も考慮しなければならない．声質の個人差の場合と同様に，このような多様な表現を収録したようなデータベース構築は難しい．

そこで，各ドラム音がどのような擬音語で表現されるかを口ドラム用の発声辞書に登録しておくことで，この問題を解決する．人によって口ドラムの表現に個人差はあるが，各ドラム音の擬音語の種類はある程度限定することができ，事前に発声辞書を構築することで表現の差異に対処できる．

人が音をどのような擬音語へ変換するのかを一般的な見地からとりあげたものに，田中ら[13),14)]，比屋根ら[15)]，石原ら[16)] による研究がある．特に比屋根ら[15)]は，単発音が擬音語へ変換される際の規則を示しており，発声辞書構築に利用できる．しかし，その規則は単発音に関するものであり，ドラムパターンを歌った口ドラムはそのような一般規則では網羅できないことも考えられる．そこで，データ収集を兼ねて，人がドラムパターンをどのような擬音語で表現するかを調査する実験も行い，その結果も発声辞書に含める．

以下では「口ドラムの表現」という語を「どのような擬音語を用いるか」という意味で用いる．

2.2 口ドラム認識の概要

本論文では，ドラム音として最も重要な Bass Drum（BD）と Snare Drum（SD）を対象として，口ドラム認識を行う．口ドラムをドラムパターンとして認識するために，本手法は，個々の楽器名と発音開始時刻を決定してからドラムパターンを決定するのではなく，入力口ドラムがデータベース中のどのドラムパターンに近いかを直接求める．このような手法を採用することで，認識における楽器名の誤挿入などの補正を必要とせず，既存ドラムパターンを得ることができる．

具体的には，2 段階の処理によって口ドラムを認識する．まずドラムパターンの擬音語系列を用いて，入力口ドラムの楽器名の並び（e.g., BD SD BD BD SD. 以下，**シーケンス**と呼ぶ）を決定する．次に，認識されたシーケンスに対して発音間隔（IOI: Inter-Onset Interval）を算出し，それが最も類似しているパターンを最終的な認識結果として出力する．このように 2 段階で処理を行うため，正解ドラムパターンとして認識されるためには，第 1 段階の処理でシーケンスが正しく推定されている必要がある．

データベース登録パターンを出力とする本手法の長所としては，ドラム演奏や口ドラム発声に熟達していないユーザの入力を自動補正（クオンタイズ）できることがあげられる．すなわち，あるドラムパターンを得たいが，（意図的なずらしを含めて）完璧には発声できないユーザに対して特に有用である．現実に使用されるドラムパターンは（メロディラインなどに比べると）それほど多くの種類があるわけではないため☆，既存ドラムパターンの出力でそのようなユーザの要求を十分カバーできると考える．ドラムパターンデータベースを拡張することで，対応範囲を広げることも可能である．さらに，ドラム演奏や口ドラム発声に熟達したユーザが，口ドラムの発音開始時刻や音量をより反映させた出力結果を得たいような状況では，第 1 段階で決定されたシーケンスに対し，発音開始時刻と音量を推定してそれを利用することも考えられる．

2.3 処理の流れ

認識処理の具体的な流れを**図 1** に示す．システムはドラムパターンデータベース（図 1，D），発声辞書（B2），音響モデル（B4）から構成される．入力された口ドラム音（A1）に対し，音声認識で広く用いられている音響特徴量 MFCC（Mel Frequency Cepstral Coefficients），ΔMFCC，ΔPower を求め（A2），全ドラムパターンの各擬音語系列を表現したネットワーク（B5，擬音語の音響モデルを連結したネットワーク）との尤度を Viterbi アルゴリズムによって計算する（A3）．その結果，各シーケンスの尤度とドラム単音ごとの発音開始時刻が得られるので，その尤度が最も高いものをシーケンスの認識結果とする（A4）．ただし，発音開始時刻には，パワーが大きく安定した結果が得られる母音の開始時刻を用いる．これらの実装のうち，音響特徴量の抽出（A2）と尤度計算（A3）には，HTK Speech Recognition Toolkit[18)] を利用

☆ 4 章で後述するように，RWC 研究用音楽データベース（ポピュラー音楽）[17)] 90 曲から得られたドラムパターンは 1,230 種類．

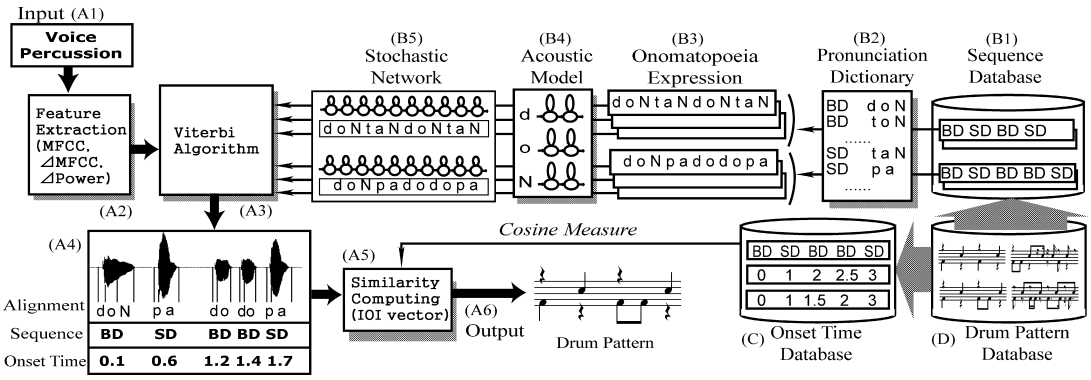

図 1 手法の流れ（A2 と A3 の実装には HTK Speech Recognition Toolkit [18] を用いた）
Fig. 1 Overview of the system (HTK Speech Recognition Toolkit [18] used in A2 and A3).

した．

第 2 段階の処理は，（A4）の認識結果とドラムパターンデータベース中の同一シーケンスの全パターンとの発音開始時刻系列の類似度を求めることによって（A5），最終的な出力（A6）を得る．類似度計算は，各発音開始時刻の系列から IOI をベクトルとして求め，ベクトル間のコサイン尺度値を算出することによって行う．発音開始時刻を並べたベクトル $\boldsymbol{o}$ に対し，IOI ベクトル $\boldsymbol{d}$ は $\boldsymbol{o}$ の差分ベクトルで，発音開始時刻の間隔を表現するために広く用いられている手法である．図 1 の例では，$\boldsymbol{o} = (0.1, 0.6, 1.2, 1.4, 1.7)$ に対し，$\boldsymbol{d} = (0.5, 0.6, 0.2, 0.3)$ である．

2.4 発声辞書の構築

発声辞書を構築するためには，ドラム音がどのような擬音語として発声されるのかを知る必要がある．比屋根らは，短時間に減衰する単発音に対して，中心周波数や残響時間，周波数ゆらぎなどのパラメータによって 6 種類の擬音語の分類を報告している[15]．そこで提案された分類においては「衝突減衰音」と「多重衝突音」がドラム音にあてはまる．衝突減衰音はある物体間の 1 回の衝突がごく短時間に減衰する音を示し，多重衝突音は同系の音が 2 回以上連続して発音される音を示す．

BD の擬音語には，衝突減衰音として「**ト**」「**トッ**」「**トン**」「**ド**」「**ドッ**」「**ドン**」「**ズ**」「**ズッ**」「**ズン**」を用いた．SD の擬音語には，衝突減衰音として「**タ**」「**タッ**」「**タン**」「**パ**」「**パッ**」「**パン**」を用いた．また，多重衝突音も考慮し，BD に「**コ**」（トコトコ，ドコドコ のため），SD に「**カ**」（タカタカ のため）も含めた．

このように構築した辞書を基本的な発声辞書とする．また，口ドラム歌唱の実験を行い（3 章で後述），そこで用いられた擬音語も発声辞書に追加登録した．

2.5 音響モデルの口ドラムへの適応

口ドラム音は，通常の会話では使わない特殊な音声であるため，音声認識で用いられる日本語音素の音響モデルをそのまま用いると，音響的な特性の違いにより誤認識する可能性が大きい．そこで，音響モデルを口ドラム音へ適応させることを考える．これには，音声認識において一般的に用いられる話者適応の手法[19]を適用することができる．そこで本実験では「口ドラム音（発声スタイル）への適応」と「口ドラム歌唱者（話者性）への適応」の 2 種類の適応を行う．口ドラム音への適応では，歌唱者以外の口ドラム音を適応データとして用い，口ドラム歌唱者への適応には，適応データにその歌唱者自身の口ドラム音を用いる．前者は汎用的な状況，後者は歌唱者を限定できる状況での利用を想定している．

音響モデル適応の際のパラメータ推定手法として MLLR（Maximum Likelihood Linear Regression）と MAP 推定（Maximum *A Posteriori* Probability）を組み合わせた MLLR-MAP を用いた[20]．MLLR は，モデルパラメータ間での情報の共有化を利用した方法である．MAP 推定は適応学習における事前知識を効率的に利用した方法であり，推定に用いる観測データが少量の場合でも頑健なモデル推定が行えることが知られている．MLLR-MAP は，まず MLLR によってモデルパラメータの変換を行い，それを事前知識として MAP 推定を行う．MLLR-MAP を用いることにより，MLLR，MAP 推定それぞれ単独で適応を行う場合と比べて，高精度に適応可能であることが報告されている[20]．本研究では，MLLR-MAP による音響モデルの適応には HTK Speech Recognition Toolkit [18] を用いた．

3. 口ドラム表現実験

ドラムパターンの口ドラム表現を，心理実験により調査する．本実験では，19 歳から 31 歳の男女 17 人の被験者（打楽器演奏経験者 2 人，非経験者 15 人）による口ドラムの歌唱を収録して分析した．対象とするドラムパターンは，Bass Drum（BD）と Snare Drum（SD）のみで構成され，音は 2 つ同時に鳴らないものとし，4/4 拍子の 1 小節とした．

各被験者は，BD と SD のみで構成されるドラムパターンを聴取し，それに対応する擬音語をイメージして歌唱する．BD と SD の音源には，後藤らの開発した RWC 研究用音楽データベース（楽器音）[17] に収録されている RWC-MDB-I-2001 No.42「ロックドラムス 1」の強さ「強」のデータを用いた（421BD1N3.WAV, 421SD3N3.WAV）．

3.1 実験方法

各被験者は，実際のドラム音で演奏されたドラムパターンの再生音を，その楽譜も見ながら，記憶するまで何度も聴取する．その後，再生音を停止し，被験者は自分で擬音語表現を考えて歌唱する．本実験を通じて，被験者には口ドラム歌唱の例は示さない．

実験に用いたドラムパターンは，ドラムの練習教本を参考に決定した 10 パターンを用いた（**図 2**）．各パターンに対して，M.M. = 80 と M.M. = 120☆の 2 種類のテンポで演奏した再生音を用意し，被験者 1 人あたり 20 パターンの口ドラムの歌唱を収録した．ドラムパターンの呈示順は，被験者ごとにランダムとした．

3.2 実験結果

被験者には，聴こえたパターンを擬音語に直して発声するように指示したが，BD と SD の違いを，同一擬音語の音高の違いのみを用いて表現する被験者も見られた．そのうちの 1 人は全パターンを音高の違いで表現し，3 人は部分的に音高の違いで BD と SD を区別していた．後者については，呈示ドラムパターンが速い場合や複雑である場合に，このような傾向が見られた．

擬音語の違いでドラム音を表現した場合，その表現の仕方は，“CV(Q)”，“CVN(Q)”，“CVRN(Q)” の 3 つの構造のいずれかの形をとった．ここで，CV は「子音 + 母音」，Q は促音，R は長音，N は撥音を表す．たとえば，SD の CV を「タ」とすると，CVQ が「タッ」，CVN が「タン」，CVRN が「ターン」となる．CV の表現の種類を**表 1** に示す（数字はその擬音語を発声した人数☆☆）．また，ここで得られた結果を考慮した最終的な発声辞書を**表 2** に示す．

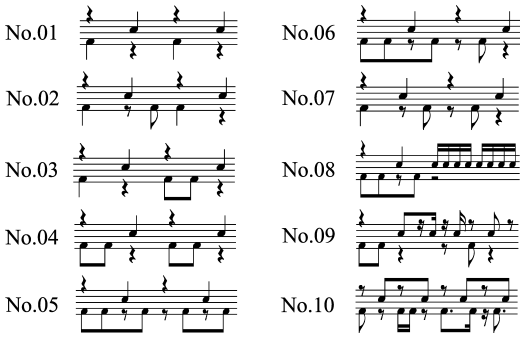

図 2 実験で用いた呈示ドラムパターン
Fig. 2 Drum patterns used in the experiments.

表 1 表現された擬音語の CV 部分の種類
Table 1 Examples of CV pairs used for drum sounds.

	BD						
CV の種類	ド	ドゥ	ト	トゥ	ズ	ク	レ
発声した人数	9	8	5	4	3	1	1
	SD						
CV の種類	タ	ダ	パ	カ	チャ	テ	ラ
発声した人数	13	3	3	2	1	1	1

表 2 発声辞書
Table 2 Pronunciation Dictionary.

	衝突減衰音	多重衝突音
Bass Drum	ト，トン，ズ，ズン，ド，ドン トーン，ズーン，ドーン ドゥ，ドゥン，ドゥーン トゥ，トゥン，トゥーン クン，レ	コ
Snare Drum	タ，タン，パ，パン ターン，パーン ダ，ダン，ダーン チャ，テ	カ，ラ

※最後に促音（Q, ッ）をつけた表現を除いて表記．

被験者の発声にはさらに，以下のような傾向が見られた．

- 有声音として発声する場合と無声音で発声する場合がある．
- 次の音までの時間的な間隔が広くなるに従って，“CV(Q)”，“CVN(Q)”，“CVRN(Q)” となる．
- BD の “CV” を「ド」と発声する被験者でも，BD の速い連続音では「ドド」ではなく「ドト」と発声する．
- 最初は音高の違いで表現していた被験者も，途中

☆ M.M.: Mälzel’s Metronome の略．メトロノームによって計られる 1 分間の拍数を表す．

☆☆ 1 人の歌唱者が複数の CV を使用することもある．

で擬音語の違いで表現し始めると，それ以後は擬音語の違いによる表現のみとなる．

ドラム・パーカッション経験者の 2 人のみ，ドラムパターン No.08 の SD の多重衝突音を「タカタカ」と発声していた（それ以外の被験者は「タタタタ」と発声していた）．また，この 2 人は休符（RE）も発声していた．たとえば，No.09 のドラムパターン「BD BD SD RE SD RE SD BD SD」に対して「ズンズンタンズタズタズッタン」と，BD と同じ擬音語を用いたり「ドンドンタンタンタドンタン」のように「ン」を用いた発声を行っていた．

3.3 考　　察

本実験の結果から，口ドラム認識のための発声辞書が構築でき，ドラムパターンにおける口ドラム表現の知見が得られた．口ドラム表現では，音色を表現するために音高を用いる場合があること，無声音のような特殊な発声をする場合があること，ドラムパターンを擬音語に変換する場合，単発音の擬音語変換とは違う規則が存在することが分かった．以下，それぞれの知見への対処について考察した後，本手法の有効範囲を述べる．

BD と SD を音高の違いで表現

音高の違いで表現することは例外的として扱い，現段階では対処しない．こうした表現を被験者がした理由には，以下の 3 つが考えられる．

- 複雑なパターンや速いパターンを呈示された場合パターンを追うことで精一杯になり，言葉を使い分ける余裕がなかった．
- 初めだけ音高の違いで表現していた被験者
  各ドラム音をどのような擬音語表現で歌うとよいか，初めはうまく考えられていなかった．
- すべてを音高の違いで表現していた被験者
  音高の違いのみで十分 BD と SD の違いを表現できると考えていた．

すなわち，音高も考慮に入れることで，認識性能が向上すると考えられるため，将来はこのような表現にも対処する必要がある．また，認識対象のドラム音を増やす場合に，同一擬音語が用いられる異なるドラム音に対処できる可能性もある．

口ドラム特有の発声・擬音語

本手法では，口ドラム特有の発声を，音響モデルの適応によって対処する．無声音で発声される口ドラムは，通常の会話と異なる特殊な発声の代表的な例である．無声発声を行った被験者は，ドラム音を擬音語として表現し，かつ，実際のドラム音を音響的に模倣しようとしていると考えられる．特に，ドラム演奏経験者が顕著に無声発声をしていた．また，声と息を同時に出すような発声（中国語における有気音のような発声）をした被験者もいた．

ドラムパターンを歌った口ドラム特有の単発音の擬音語変換と異なる規則については，発声辞書を再構築することで対処した．

多重衝突音の擬音語，休符の発声

本手法では多重衝突音の擬音語には対処するが，休符の発声には対処していない．これらの現象は，楽器演奏の練習方法に起因すると考える．たとえば，ドラム教則本には，リズムのとり方を理解させるために，「タカタカ」や「ンタンタ」などの記載がある場合が多い．打楽器以外の楽器の演奏者でも，休符を読んだり，多重衝突音の擬音語を発声したりすることでリズムをとりやすくした経験を持っている可能性がある．将来は，休符の発声にも対処する必要がある．

提案手法の有効範囲

本実験において，現段階では対処できない「音高の違いによる表現」と「休符の発声」の割合は，被験者 17 人 20 パターンの 340 発話中，それぞれ 9.11%（31/340 発話）と 0.88%（3/340 発話）であった．すなわち，提案手法は 90%（306/340 発話）の口ドラムに対して適用可能である．

4. 口ドラム認識実験

口ドラム表現実験で収録した口ドラム音を利用して，本手法で正しいドラムパターンとして認識できるかを評価する．認識のためのドラムパターンデータベースは，RWC 研究用音楽データベース（ポピュラー音楽：RWC-MDB-P-2001）[17] のドラムスを含む 90 曲の SMF（Standard MIDI File）から，ドラムトラックの各小節を機械的に切り出して構築した．口ドラム表現実験と同様に，BD と SD の発音時刻が同一であった場合は SD を優先させて同時に 1 音のみが鳴るものとし，4/4 拍子の区間のみを利用した．また，口ドラム表現実験で用いたドラムパターンは，データベースとは無関係に決定したため，そこで用いられたパターンもデータベースに含めた．このようにして，592 種類の異なるシーケンスからなる，全 1,230 種類のドラムパターンを登録した☆．

本実験では，異なる音響モデルと発声辞書を組み合わせた，以下のような 4 種類の実験条件で口ドラムの認識率を評価した．

☆ ドラムパターンの種類が多いのは，楽器名の並び（シーケンス）が同一でも発音時刻が異なるパターンがあるためである．

(A) 通常の音声認識用の音響モデルを用いる．

(B) 評価用データ（後述）に含まれない口ドラム音で適応させた音響モデルを用いる．

(C) 各被験者ごとに話者適応を行った音響モデルを用いる．

(D) 各被験者ごとに話者適応を行った音響モデルを用い，さらに各被験者が発声した表現だけを登録した発声辞書を用いる．

口ドラム表現実験で収録した 17 人の口ドラムデータのうち，評価用データとして 10 人 200 発話☆，条件 (B) の適応用データとして，評価用データには含まれない 5 人 100 発話を用いた．17 人の残り 2 人は，ほとんどすべてのパターンに対して，BD と SD の違いを同一擬音語の音高の違いのみで表現したため，評価対象から除外した．ただし，評価用 200 発話には，現段階では対処していない，音高の違いのみを用いたデータが計 3 発話（= 1.5%）と，休符の発声されたデータが計 3 発話（= 1.5%）含まれている．

条件 (C)，(D) は，認識率を cross validation 法で評価した．具体的には，各被験者の発声 20 パターンを適応用 18 パターンと評価用 2 パターン（テンポの異なる同一パターンの組）に分け，10 種類のドラムパターンそれぞれを「評価用」とした 10 回の実験を行い，その認識率で評価した．

4.1 評価方法

パターン評価と発音時刻評価の 2 種類の評価を行う．

パターン評価は，認識の結果得られたシーケンス（図 1，A4）とドラムパターン（A6）が，それぞれ正しく推定されたかどうかを評価する．シーケンスでは楽器名（BD/SD）の系列，ドラムパターンでは楽器名と発音開始時刻の系列が完全に一致していた場合のみを正解とする．すなわち，入力口ドラムに対し，正解シーケンスが尤度最大であった場合，発音開始時刻から得られる IOI ベクトルが正解ドラムパターンと最も類似していた場合がそれぞれ正解となる．認識率は，評価用口ドラム 10 人 200 発話において，以下のように定義する．

$$\text{認識率} = \frac{\text{正しく認識した発話数}}{\text{総発話数}} \times 100$$

また，ドラムパターンが正解となるにはシーケンスが正解である必要があるため，シーケンスの認識率がドラムパターンの認識率の上限となる．

発音時刻評価は，各口ドラム音声に対して正解時刻のラベル付けを人手で行い，その時刻と，認識された発音開始時刻（図 1，A4）との「ずれ」を評価する．発音時刻評価は，条件 (D) において，パターン評価でシーケンスを正しく推定できた発話について評価した．

4.2 実験結果

評価用 200 発話を入力（図 1，A1）とし，592 種類のシーケンス（B1）からなる，全 1,230 種類のドラムパターンで構築されたデータベース (D) を用いて認識を行った結果を評価する．発声辞書における口ドラム表現（擬音語）の登録数は，共通辞書で 62，個別辞書で平均 3.83 であった．

(1) パターン評価

表 3 にシーケンスおよびドラムパターンの認識率，**表 4** にシーケンスの N-best 認識率を示す．N-best 認識率は，尤度が上位 N 位までの結果に正解が含まれる場合の認識率であり，$N = 1$ の結果は表 3 に対応する．

図 3 に，被験者ごとの認識率のグラフを示す．縦軸が認識率，横軸は被験者番号であり，性別による認識

表 3 認識実験の各条件と結果
Table 3 Experimental conditions and results.

条件	音響モデル	発声辞書	認識率	
			Sequence	Drum Pattern
(A)	適応なし	共通辞書	61.0%	58.5%
(B)	歌唱者非依存	共通辞書	60.0%	58.5%
(C)	歌唱者依存	共通辞書	86.5%	85.0%
(D)	歌唱者依存	個別辞書	93.5%	92.0%

表 4 シーケンスの N-best 認識率
Table 4 Recognition rates of N-best sequences.

条件	認識率			
	$N = 1$	$N = 3$	$N = 5$	$N = 10$
(A)	61.0%	71.0%	73.5%	78.0%
(B)	60.0%	73.5%	78.5%	85.0%
(C)	86.5%	94.0%	95.0%	96.0%
(D)	93.5%	95.5%	96.0%	96.5%

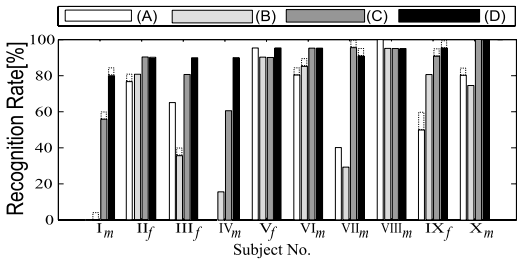

図 3 パターン評価の結果（被験者ごと認識率）．被験者番号の添え字は性別を示す（m = 男，f = 女）

Fig. 3 Result of pattern evaluation (by subject). The subscript of subject No. means subject's gender (m = male, f = female).

☆ 図 2 の 10 パターンを 2 種類のテンポで歌った 20 発話ずつ．

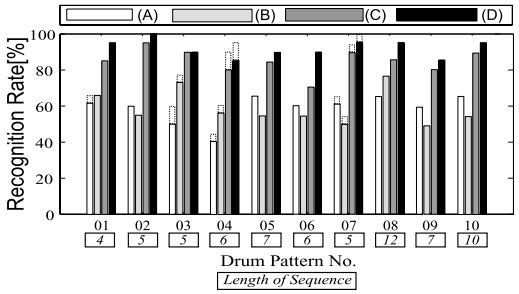

図 4 パターン評価の結果（ドラムパターンごと認識率）．$\boxed{n}$ のように四角で囲まれた数字は，そのシーケンスの長さを示す
Fig. 4 Result of pattern evaluation (by drum pattern). The length of sequence is shown in boxed form, like $\boxed{n}$.

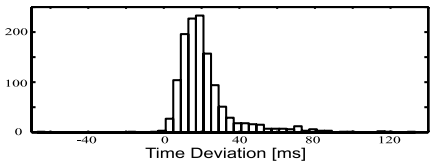

図 5 発音時刻評価の結果（条件 (D) でのずれの頻度分布）
Fig. 5 Result of onset evaluation.

率の違いを考察するために被験者番号には性別を添えて表示した．ここで，被験者 I と VI が打楽器演奏経験者である．**図 4** には，ドラムパターンごとの認識率のグラフを示す．横軸は被験者が歌った図 2 におけるドラムパターン番号を示す．入力となるシーケンスの長さ n が認識率に及ぼす影響を調査する目的で，$\boxed{n}$ のように表したシーケンス長も表示した．また，図 3，図 4 の破線はシーケンス認識率，実線はドラムパターン認識率を表す．たとえば，図 3 の被験者 IX の条件 (D) では，シーケンスの認識率が 100%であったが，ドラムパターンの認識率は 95.0%であった．

(2) 発音時刻評価

図 5 に認識された発音開始時刻のずれの頻度分布を示す．横軸は，認識された発音時刻から正解発音時刻を引いた値である．ずれの平均は +21.3 ms，標準偏差は 15.5 ms，最も大きなずれ幅は +133.0 ms であった．

4.3 考　　察

パターン評価において，条件 (A) では，日本語用音素の音響モデルを用いているために，口ドラム音の特殊な発声である無声発声に対して誤認識が頻出した (e.g., 被験者 I)．また，条件 (B) では，無声発声の認識率があまり改善されなかったこと（図 3，被験者 I），認識率が低下してしまう例（被験者 III）が存在したことにより，平均認識率が条件 (A) に比べてほとんど変化しなかった（表 3）．これは，適応用データに被験者と同じ発声スタイルのデータが少なかったことが原因の 1 つとして考えられる．最も認識率が低かった被験者 I は BD を「ズッ」(/zu/よりは/z/のような音）と発声し，そのような発声をした被験者はほかにいなかった．しかし，認識率が大きく向上する例（被験者 IX）も見られたこと，正解シーケンスの尤度は条件 (A) に比べて上位となっていたこと（表 4）から，口ドラム特有の音響的性質への適応はできていると考えられる．すなわち，適応用データを発声スタイルに合わせてうまく選別するような手法が導入できれば，さらに良い認識率を得られる可能性がある．条件 (C) では，無声音の発声に対する誤認識がかなり減少し，個別辞書を用いた条件 (D) では，ほぼ誤認識しなくなる．条件 (D) でも誤認識してしまうケースとしては，「休符を発声する」，「同一擬音語を用いて音高の違いのみでドラム音の違いを表現する」が原因であった．

また，本手法における認識性能は，上述のように，発声スタイルに大きく依存し，歌唱者の性別には依存していなかった．たとえば，男性被験者 VIII と女性被験者 V はどちらも，すべての条件下で高い認識率を得ていた（図 3）．また，シーケンスの長さによる認識率の差についても，目立った違いはなかった（図 4）．

発音時刻評価においては，ずれの標準偏差 15.5 ms が，M.M. = 120 における 64 分音符の長さ（31.25 ms）よりも小さな値であり，十分な精度で認識できているといえる．発音開始時刻として母音の開始を採用しているためにずれの平均は正（+21.3 ms）だが，類似度計算には IOI を用いるので，認識には影響しない．

本実験で得られた結果は，本手法が高い有効性を持つことを示している．不特定多数の歌唱者に対して運用する場合，条件 (B) のように口ドラム音声を用いて既存の音響モデルに適応処理を行うことで，より良い認識率を得ることができる．さらに応用システムでは，適応用データの選別を行わない場合でも，インタフェースを工夫し，5-best を利用するような運用をすれば，80%弱の精度で正解を得ることができる．ここで，無声音を使用しないなど，発声スタイルを制限することでさらに良い認識率が得られる．また，歌唱者の口ドラムデータを事前に収集可能な状況であれば，条件 (C) の結果に示されたように，18 パターン程度（約 1 分）のドラムパターンを発声することで，認識率が 25%程度向上することが期待できる．さらに，歌唱者が自分の口ドラムで用いる擬音語表現も事前に登録すれば，認識率は 90%を超え，実用性はさらに高くなる．

5. ドラム譜入力インタフェース：Voice Drummer

提案手法により口ドラム認識が実用可能と考えられることを受けて，口ドラムでドラムパターンを入力することができる新しい楽譜入力インタフェース Voice Drummer を実装した．実装には，Microsoft Visual C++ .NET 2003 を使用した．**図 6** にその画面例を示す．

5.1 Voice Drummer の機能

口ドラムによる入力が作曲や編曲においても有効であることを実証するために，作曲を想定したドラムパターンの入力機能（**楽譜入力モード**）と既存の楽曲のドラムパートだけを差し替えて編曲する機能（**編曲モード**），さらに，ユーザが練習しながら自分の声を学習させることで口ドラム入力の認識率を上げることができる歌唱者適応機能（**練習適応モード**）の 3 つの機能を用意した．図 6 の表示画面例の右下の “Notation”，“Arrange”，“Practice” のボタン（図 6，A）を押すことで，上記の各機能を呼び出すことができる．

楽譜入力モードと編曲モードでは，4 章で構築したドラムパターンデータベース（1,230 種類）を用いて口ドラム認識を行い，小節単位で入力する．尤度計算（図 1，A2）の実装には Julian [21] を用い，特徴抽出（A3）と練習適応モードにおける適応は HTK Speech Recognition Toolkit [18] で実装した．

表示画面の上部は，以下の 3 つのウィンドウで構成される．

- 楽譜ウィンドウ（B，“Drum Score”）
  口ドラムの認識結果などによるドラムパターンが，4 小節 × 2 段の構成で 8 小節分表示される．各小節内は 2 段に分かれ，下段の青い（図 6 では黒）長方形のマークは Bass Drum，上段の緑の（図 6 では灰色）長方形のマークは Snare Drum の発音時刻をそれぞれ表す．
- ストリームウィンドウ（C，“Drum Stream”）
  中央の縦長のバー（E）が現在時刻を表し，入力対象のドラムパターンがその右から左へと流れてゆく．3 段に分かれ，下 2 段は楽譜ウィンドウと同じように Bass Drum と Snare Drum が表示され，Hi-hat の発音時刻を表すマークが表示される．
- 口ドラムウィンドウ（D，“Voice Percussion”）
  中央の縦長のバー（F）の中に，マイクから入力された声の大きさ（パワー）が表示される．このパワー値表示は，時間とともに左へと流れてゆく線グラフで表示され，ユーザの発声音量が適切であるかどうかが確認できる．中央のバーの右側には，練習適応モードにおけるユーザへの適応の度合いを示すグラフ（G）が表示される．

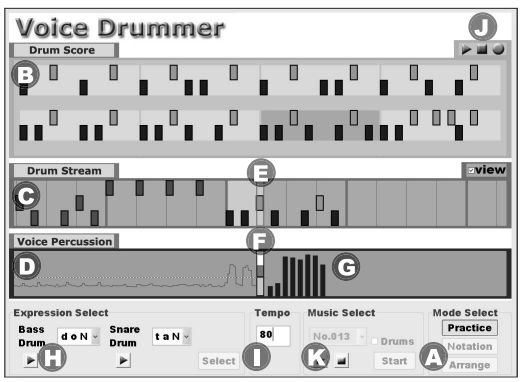

図 6 Voice Drummer の表示画面（練習適応モード）
Fig. 6 An example Voice Drummer screen (practice/adaptation mode).

ユーザは，入力状況や認識結果がリアルタイムに反映されるこれらのフィードバック画面を見ながら，まず，練習適応モードで練習し，その後，他の 2 つのモードでドラム譜を入力することができる．

5.1.1 擬音語の決定

Voice Drummer は，Bass Drum と Snare Drum を対象とし，それらの多様な擬音語表現に対応する．発声辞書は，3 章で作成したものを用いる．ただし，認識精度や反応速度を向上させるために，図 6 下部の “Expression Select” のメニュー（H）により，認識対象とする表現をユーザが限定することも可能とした．

5.1.2 練習適応モード

ユーザは，楽譜ウィンドウに提示される 8 種類のドラムパターンの中から 1 つ（1 小節）を選択して，その練習を行うという操作を繰り返す．まず，ストリームウィンドウ上を，選択したドラムパターンが右から流れてくる．中央の現在時刻のバーに到着すると，図 6 下部の “Tempo”(I) で指定したテンポでドラム音が MIDI 音源で演奏される．発声するタイミングをつかみやすいように，各ドラムパターンに先だって，4 分音符で 4 回（1 小節分）Hi-hat 音が演奏される．次に，同一パターンが再度右から流れてくるので，それが現在時刻のバーに重なるタイミングで口ドラムを歌唱する．

これはユーザの練習になるだけでなく，そのユーザの声を学習することで口ドラム認識率を上げるためにも有効である．これによりユーザが楽しみながら適応もできることを目指す．n 回目の練習を終えると，図 6

に示すように，口ドラムウィンドウの右側にユーザへの適応度を反映した棒グラフが $n-1$ 本表示される（G）．このグラフは，n 回目の口ドラムに対する適応後の尤度の上昇率を示しており，練習終了の目安として用いることができる．

5.1.3 楽譜入力モード

ユーザが 1 小節のドラムパターンを口ドラムで歌唱すると，認識された結果が MIDI 音源で演奏され，楽譜ウィンドウに 1 小節ずつ順番に表示される．こうして作成したドラム譜はボタンを押すことで再生でき，標準 MIDI ファイル（SMF）への保存ができる（J）．

さらに，ユーザの発想支援やドラムパターンの知識を増やす目的で，ストリームウィンドウに，データベース中のドラムパターンを一定の頻度でランダムに流す機能もある．

5.1.4 編曲モード

Bass Drum と Snare Drum の音が消去された楽曲が 8 小節分演奏されるので，ユーザはこれに合わせて編曲したいイメージどおりに口ドラムの歌唱を行う．口ドラムは 1 小節ごとに次々と認識され，楽譜ウィンドウに順次入力される．こうしてドラムパートを編曲した楽曲は，ボタンを押すことで再生することができ，SMF にも保存できる（J）．なお，練習適応モードのように，楽曲に先だって 1 小節分の Hi-hat が流れる．

対象曲は図 6 下部の “Music Select” のメニュー（K）で選曲する．“Drums” チェックボックスにチェックを入れることで，原曲がドラムパート付きで再生される．原曲のドラムパートを参考に口ドラムを歌いたい場合には，ストリームウィンドウに原曲のドラムパターンを流すこともできる．

5.2 Voice Drummer の有用性

Voice Drummer を実際に運用した結果「ドラムを叩いたことはないが，簡単にドラムパターンが入力できた」，「作曲に使いたい」などのコメントを得ることができた．また，ドラムスの演奏経験があるユーザから「ドラムの練習のために口ドラムを歌うことがあるので，練習に役立ちそう」というコメントも得ることができた．さらに詳細な評価についても，今後行う必要がある．

6. おわりに

本研究では，表現の個人差にも対応できる口ドラム認識手法を実現した．これによって，音声（歌唱）によるドラムパターンの入力が可能となっただけでなく，口ドラムを擬音語で表現することで不特定歌唱者に対応できることが明らかになった．本手法をシステムとして実装する際には，擬音語を特定して 1 分程度の口ドラム発声により歌唱者への適応を行うことで，十分実用的な精度が得られる．

さらに，口ドラムによるドラム譜入力インタフェース Voice Drummer を実装した．Voice Drummer は，口ドラムの練習と同時にシステムを歌唱者に適応させることができるため，不特定歌唱者への対応を視野に入れたアプリケーションとして有用である．また，ドラムスの演奏経験がなくても，ドラムパートの作曲・編曲が容易に行えるようになった．

今後は，Bass Drum と Snare Drum 以外のパーカッション（Hi-hat，Tomtom，Cymbal）の口ドラムも認識可能とすることを目標とし，楽曲検索のためのアプリケーションの構築を目指す．このためには，それぞれの擬音語表現の調査，同一擬音語表現への対処，同時発音への対処，休符への対処，などが必要となる．また，Voice Drummer では「練習」としての適応インタフェースを実現したが，今後は，ゲーム性を取り入れた「楽しい」適応に関する研究も進めたい．

謝辞 本研究の口ドラム表現実験に関してご助言をしていただいた山本那美氏，口ドラム表現実験に参加していただいた被験者の方々に感謝いたします．本研究では，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001，楽器音 RWC-MDB-I-2001）を使用しました．

参考文献

1) 蔭山哲也，高島洋典：ハミング歌唱を手掛りとするメロディ検索，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1543–1551 (1994).
2) 園田智也，後藤真孝，村岡洋一：WWW 上での歌声による曲検索システム，電子情報通信学会論文誌，Vol.J82-D-II, No.4, pp.721–731 (1999).
3) 小杉尚子，櫻井保志，山室雅司，串間和彦：Sound Compass：ハミングによる音楽検索システム，情報処理学会論文誌，Vol.45, No.1, pp.333–345 (2004).
4) Herrera, P., Yeterian, A. and Gouyon, F.: Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques, *Proc. International Conference on Music and Artificial Intelligence* (*ICMAI*), LNAI2445, pp.69–80 (2002).
5) 後藤真孝，村岡洋一：打楽器音を対象にした音源分離システム，電子情報通信学会論文誌，Vol.J77-D-II, No.5, pp.901–911 (1994).
6) Zils, A., Pachet, F., Olivier, D. and Gouyon, F.: Automatic Extraction of Drum Tracks from Polyphonic Music Signals, *Proc. WEb DELiv-*

ering of Music (*WEDELMUSIC*), pp.179–183 (2002).
7) 吉井和佳，後藤真孝，奥乃　博：テンプレート適応を利用した実世界の音楽音響信号に対するドラムスの音源同定，情報処理学会研究報告音楽情報科学 2003-MUS-53-12, Vol.2003, No.127, pp.55–60 (2003).
8) Kapur, A., Benning, M. and Tzanetakis, G.: Query-by-Beat-Boxing: Music Retrieval for the DJ, *Proc. 5th International Conference on Music Information Retrieval* (*ISMIR2004*), pp.170–177 (2004).
9) Hazan, A.: Towards Automatic Transcription of Expressive Oral Percussive Performances, *Proc. International Conference on Intelligent User Interfaces* (*IUI'05*), pp.296–298 (2005).
10) 中野倫靖，緒方　淳，後藤真孝，平賀　譲：口ドラムによるドラムパターン検索手法，情報処理学会研究報告音楽情報科学 2004-MUS-55-8, Vol.2004, No.41, pp.45–50 (2004).
11) Gillet, O. and Richard, G.: Drum loops retrieval from spoken queries, *Journal of Intelligent Information Systems*, Vol.24, No.2–3, pp.159–177 (2005).
12) 河原達也，住吉貴志，李　晃伸，坂野秀樹，武田一哉，三村正人，伊藤克亘，伊藤彰則，鹿野清宏：連続音声認識コンソーシアム 2002 年度版ソフトウェアの概要，情報処理学会研究報告音声言語情報処理 2001-SLP-48-1, Vol.2003, No.48, pp.1–6 (2003).
13) 田中基八郎，松原謙一郎，佐藤太一：異音の表現における擬音語の検討：衝突音等の単発音やうなり音の場合，日本機械学会論文集 C 編，Vol.61, No.592, pp.4730–4735 (1995).
14) 田中基八郎，松原謙一郎，佐藤太一：機械の異常音の擬音語表現，日本音響学会誌，Vol.53, No.6, pp.472–482 (1997).
15) 比屋根一雄，澤部直太，飯尾　淳：単発音のスペクトル構造とその擬音語表現に関する検討，電子情報通信学会研究報告 SP97-125, pp.65–72 (1998).
16) 石原一志，坪田　康，奥乃　博：日本語の音節構造に着目した環境音の擬音語への変換，電子情報通信学会研究報告 SP2003-38, pp.19–24 (2003).
17) 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).
18) Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P.: *The HTK Book*, Version 3.2.1, p.346 (2002).
19) 篠田浩一：確率モデルによる音声認識のための話者適応化技術，電子情報通信学会論文誌，Vol.J87-D-II, No.2, pp.371–386 (2004).
20) Digalakis, V.V. and Neumeyer, L.G.: Speaker adaptation using combined transformation and Bayesian methods, *IEEE Trans.Speech and Audio Processing*, Vol.4, No.4, pp.294–300 (1996).
21) 河原達也，李　晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol.20, No.1, pp.41–49 (2005).

(平成 17 年 12 月 19 日受付)
(平成 18 年 10 月 3 日採録)

中野　倫靖（学生会員）

2003 年図書館情報大学卒業．2005 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．現在，同大学院図書館情報メディア研究科博士後期課程．日本音響学会，日本音楽知覚認知学会各会員．

緒方　　淳（正会員）

1998 年龍谷大学理工学部電子情報工学科卒業．2000 年同大学大学院修士課程修了．2003 年同大学院博士後期課程修了．同年産業技術総合研究所入所，現在に至る．博士（工学）．音声認識，音声インタフェースに関する研究に従事．2000 年度日本音響学会粟屋潔学術奨励賞，2001 年度電子情報通信学会学術奨励賞，WISS2004 ベストペーパー賞各受賞．日本音響学会，電子情報通信学会各会員．

後藤 真孝（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院理工学研究科博士後期課程修了．同年電子技術総合研究所（2001 年に独立行政法人産業技術総合研究所に改組）に入所し，現在に至る．2000 年から 2003 年まで科学技術振興事業団さきがけ研究 21「情報と知」領域研究員，2005 年から筑波大学大学院システム情報工学研究科助教授（連携大学院）を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味を持つ．1997 年情報処理学会山下記念研究賞（音楽情報科学研究会），2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会粟屋潔学術奨励賞・ポスター賞，2002 年情報処理学会山下記念研究賞（音声言語情報処理研究会），2002 年日本音楽知覚認知学会研究選奨，2003 年インタラクション 2003 ベストペーパー賞，2005 年情報処理学会論文賞等 18 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．

平賀 譲（正会員）

1983 年東京大学大学院理学系研究科博士課程中退，同年図書館情報大学助手．現在，筑波大学大学院図書館情報メディア研究科教授．日本認知科学会，日本音楽知覚認知学会，ACM 等各会員．