解　説

音楽・音声の音響信号の認識・理解研究の動向

後藤 真孝　緒方 淳

本解説論文では，近年目覚ましく研究が進展した音楽の音響信号の認識・理解と，長年の研究により性能が向上した音声の音響信号の認識・理解について，その研究の現状を紹介する．インタフェースやインタラクションに関連した学会においても，音楽・音声に関わる研究成果が数多く発表されているが，音楽・音声の分野外の人達にとって適切な文献を知ることは容易でなく，それらの認識・理解技術に関する最新の状況を知ることは難しかった．インタフェースを発想する上では，詳細な実現方法より前に，どのようなことが可能になっているのかを知ることが重要である．そこで本論文では，さまざまな研究事例を紹介すると共に，研究で活用できるツールや関連学会等の情報も紹介する．

In this survey paper, we introduce recent studies on technologies for recognizing and understanding audio signals of music and speech. Music-related studies have progressed significantly in recent years and speech-related studies have improved their performances over the years. Although a lot of research results related to music and speech have been presented at academic societies for interface and interaction, it was not easy for researchers outside of music and speech fields to find appropriate references and know state-of-the-art technologies for recognizing and understanding those signals. To conceive the idea for interfaces, it is important to know what is possible before knowing how to implement. This paper therefore focuses on giving an overview of various studies and introduces useful research tools and relevant academic societies.

1　はじめに

音楽・音声に関連したインタフェースを実現したり，インタラクションを研究する上で，音楽・音声の音響信号をコンピュータが自動的に認識・理解できる技術の重要性が高まっている．そうしたインタフェースやインタラクションでは人との関わりが不可欠であり，そのために，人が演奏した音楽，人が発声した音声を入力として，その内容に基づいた処理をする必要があるからである．そして，インターネットの普及やコンピュータの高性能化，認識・理解技術の高度化により，音楽・音声を用いた研究事例は近年増えつつある．さらに社会的にも，インターネット上の音楽・音声コンテンツが増加したため，それらを的確に扱う技術が注目されている．音楽に関しては，音楽配信の携帯型デジタル音楽プレーヤの普及に伴って，今後，あらゆる音がデジタル化され，情報通信技術によって配信・検索・共有・創作・発信されることを一般ユーザが実感として持ち始め，例えば，大量の音楽を思い通りに検索したり鑑賞したりするインタフェースが期待されている．音声に関しても，ポッドキャスト等の音声コンテンツや，音声を含む動画コンテンツの配信・共有が普及し，また，ロボットや電子ペットとの音声対話等のデモンストレーションを一般ユーザが目にする機会が増えたこともあり，音声によるインタラクション技術の高度化・高信頼化が望まれている．

そうしたインタフェースやインタラクションの基礎技術として，音楽と音声の認識・理解を考えたときには，原理的に，リアルタイム処理 (オンライン処理) が可能な技術であるかどうかを区別することが大切である．人が演奏している音楽，発声している音声をリアルタイムに認識・理解できれば，それにコンピュー

A Survey on Research for Recognizing and Understanding Audio Signals of Music and Speech.

Masataka Goto and Jun Ogata, 産業技術総合研究所, National Institute of Advanced Industrial Science and Technology (AIST).

コンピュータソフトウェア, Vol.26, No.1 (2009), pp.4–24.
[解説論文] 2008 年 7 月 8 日受付.

タが反応することで，さまざまなインタラクションが可能になる．そのためには，まだ観測していない未来の音楽・音声は利用できず，場合によっては，現時点までの認識結果に基づいて未来を予測する処理が必要となる．例えば，人の演奏に合わせてコンピュータが伴奏するシステムを実現する際には，現時点までの音楽理解結果に基づいて，その先に人がどう演奏するか (例えば，テンポを早くするか等) を予測し，それに合わせて伴奏する必要がある．音声でも，対話で相手が何を発声するかを予測する場合が相当する．一方，そうではない非リアルタイム処理 (オフライン処理) も有用な場面は多く，過去に収録された音楽・音声のアーカイブ，データベース，コーパスを認識・理解することで，多様な検索やブラウジング，情報抽出等が可能になる．例えば音楽で楽曲一曲を認識・理解することを考えると，リアルタイム処理ではその途中までの音響信号しか使用できないが，非リアルタイム処理では一曲全部の音響信号を使用できる点が大きく異なる．本解説論文では，認識・理解についてのさまざまな技術を紹介するが，予測については間接的にしか扱わない．認識・理解技術は未来の予測の土台であるものの，実際に予測するには他のさまざまな要因を考慮しなければならず，むしろインタフェースやインタラクション側の研究として取り組む必要があるからである．

本解説論文では，以下，2 章 で音楽の認識・理解を，3 章 で音声の認識・理解を述べる．それぞれが研究分野として成立している大きな研究課題であるため，インタフェースやインタラクション研究者へ向けた入門解説となるように，各研究分野内での歴史的発展については敢えて触れず,「現在の技術で何が可能なのか」を幅広く紹介することに焦点を絞る．また，それを技術的にどう実現するのかという具体的な手法についても，個々の研究事例の引用論文に譲り，ここでは割愛する．4 章 では，音楽・音声の認識・理解技術を利用するために，容易に入手可能なツールやデータベースを紹介する．最後に 5 章 で，文献調査の方法や他の解説等を紹介し，まとめを述べる．

2 音楽の認識・理解

ふだん意識することは少ないが，人間が音楽を聴いてわかることは驚くほど多い．例えば，どんなジャンルの音楽か，どんな楽器で演奏されているか，歌が含まれているかどうか，メロディーはどのように変化するか，サビ (楽曲中で一番代表的な盛り上がる主題の部分) がどこにあるか，テンポが速いか遅いか等，さまざまな情報を得ることができる．本章では，まず，2.1 節 でそうした音楽から人間が認識・理解できる内容を整理する．2.2 節 では，その人間の能力をコンピュータシステム上で実現する際に，技術的に何が難しいかを述べる．そして，2.3 節 ではこれまでのさまざまな研究により可能になってきた音楽理解の技術を紹介する．

一般に，音楽での「認識」と「理解」は厳密な使い分けがまだ確立していないが，音楽の「認識」というと，パターン認識的な識別 (違いがわかる) や，同定・照合 (どれと一緒かがわかる) を意味することが多い．一方，音楽の「理解」というと，それらも包含しつつ，さらに，楽曲中の状況や，表層に現れていない現象，演奏者・作曲者の意図を把握することまで意味することがある．そこで以下では，両者を包括して「音楽理解」と記す．

2.1 人間は音楽から何が理解できるか

音楽理解研究の究極のゴールは，人間が音楽を聴いて理解できることをすべて理解できるコンピュータシステムを実現することである．人間は音楽を聴くと，好き嫌いの印象を持ったり，さまざまな感情が引き起こされたり，関連する記憶が蘇ったりする．これらはどちらかというと主観的な理解に位置付けられる．一方，より客観的な理解としては，音楽中のメロディーやサビ，テンポ，音色等の理解が挙げられる．ここでは客観的な理解を，一曲中で刻一刻と変化する事象の理解，一曲全体に対する理解，楽曲の集合に対する理解の三つに分けて議論する．

2.1.1 一曲中で刻一刻と変化する事象の理解

音楽の専門家と非専門家ではその理解内容も異なるが，多くの人が音楽を聴きながら，メロディーや歌詞

(音楽に合わせて歌える)，サビ (盛り上がる箇所かわかる)，ビート (音楽に合わせて手拍子が打てる)，テンポ (速いか遅いかがわかる) を理解できる．より詳細なレベルでは，音楽中の主要な音に関して，発音時刻 (いつ鳴ったか)，音高 (高さ)，音量 (強さ)，音色等もわかる．特に専門家の場合には，楽器種や奏法，調や転調，コード進行までわかる場合が多い．ドラムやラテンパーカッション等の打楽器音に関しても，それらを構成するさまざまな打楽器の違いを聞き分けることができる．さらに専門家の一部は，訓練を受けることで，音の絶対的な高さがわかる絶対音感を獲得していたり，音楽を聴いて楽譜を書く採譜の技術を習得していたりする．

2.1.2 一曲全体に対する理解

楽曲を聴けば，ジャンルや雰囲気，曲調がわかる．また，他の楽曲と似ているかどうか，別の知っている曲のカバーかどうかといった，他の楽曲との関係もわかることがある．歌の場合には，歌手の性別や人数 (男か女か，一人かグループか)，歌唱力 (上手いか下手か)，声質 (太い声かハスキーか等)，歌詞の言語 (知っている言語の場合) がわかり，良く知っている歌手の場合には，その歌手名までわかることが多い．有名な演奏家や指揮者の場合にも，その名前がわかることがある．専門家の中には，楽曲の周波数特性 (イコライジング) や複数の楽器パートの音のバランス (ミキシング)，音量変化の度合い (ダイナミックレンジ，トータルコンプの効き具合[†1]) 等の，詳細な音質の違いがわかる人もいる．

2.1.3 楽曲の集合に対する理解

楽曲の集合として代表的なものに，一枚の音楽 CD やレコードに収録された楽曲群であるアルバムや，複数の楽曲の曲順を指定するプレイリストが挙げられる．通常，それらは特定の意図でまとめられており，曲順も注意深く決定されていることが多い．そうした集合ごとの音響的な性質や，曲から曲への流れ (曲調の変化等) も人間は理解することができる．さらに大きな集合としては，アーティストごとの楽曲群，音楽ジャンルのような分類がある．アーティストやジャンル同士の関係 (似ているか，影響を受けているか等) を理解したり，それらを階層的に分類・理解したりすることもある．例えば，特定のアーティストの楽曲群を前期や後期のように分けたり，クラシック音楽というジャンルの楽曲群を，その楽器構成や歌唱の有無から，交響曲やピアノ曲，歌曲のようなさまざまなカテゴリ・サブジャンルに分類したりすることもできる．

2.2 なぜ音楽理解は難しいか

音楽の非専門家が容易にわかるようなことであっても，コンピュータにそれを自動的に理解させるのは難しい．それは，音楽理解が,「複数の音が相互に関係し合いながら時間的な構造を形成して内容を伝える」信号を理解するという，従来未解決で本質的な課題を含んでいるからである．特に複数の音が同時に鳴っている混合音では，その中の個々の音を分離したり理解したりすることが難しく，まだ人間の理解能力の一部しか技術的に実現されていない．

音が一つしか鳴っていなければ，その音高や音量，音色を分析することは比較的容易である．歌声の母音やピアノ音，ギター音のような音の高さが明確な音は，基本周波数 (F0) [†2]の成分と，その整数倍の周波数の高調波成分 (倍音成分，調波成分) から成る．この構造を，高調波構造 (あるいは倍音構造，調波構造と呼ばれることもある) と呼ぶ．ここで，実際には複数の異なる周波数に成分を持っているにも拘らず，整数倍の成分はグルーピングされて，基本周波数の高さの一つの音として知覚される点が重要である．そして，同じ高さでも，音色や奏法等によって，基本周波数の時間変化や，高調波成分の音量の相対比率とその時間変化が異なる．さらに，各音の鳴り始めには，上記の高調波構造に加え，周波数が整数倍の関係にない成分 (非高調波成分，非調波成分) も含まれることが多い．例えば，ピアノ音では弦が叩かれる瞬間の

†1 トータルコンプとは,「コンプレッサ」と呼ばれるダイナミックレンジを圧縮するエフェクタを，楽曲全体に適用することを意味し，音圧感を増す効果がある．近年の日本のポピュラー音楽で多用されている．

†2 基本周波数あるいは F0 (fundamental frequency) は物理量であり，その知覚量であるピッチ (pitch) と混同されることが多い．本解説論文では，音高も基本周波数の意味で用いる．

音，ギター音では弦が弾かれる瞬間の音が相当する．歌声の子音や打楽器音 (ドラム音等) では，そうした非高調波成分だけの場合もある．このように一つの音だけでもさまざまな周波数成分を含んでいるが，音が一つしか鳴っていないと事前にわかっていれば，すべての成分はその音のものだとみなして分析できる．

一方，混合音の場合には，各周波数成分がどの音に起因するものかを推定する必要があるだけでなく，異なる音の周波数成分が同じ高さで重なることもあるため，その理解は難しくなる．1 オクターブ離れた二つの音が同時に鳴っていると，低い音の基本周波数の偶数倍の成分は，高い音の高調波成分と重なることになる．楽曲中でさまざまな音が調和して響くためには，このような重なりが頻繁に起きることが必要であり，より一層問題を難しくしている．これを個々の音へ分離する「音源分離」問題は，適切な制約や仮定，知識無しには解くことはできない．また，重なったまま理解しようとしても[†3]，各音の周波数成分は他の音の周波数成分と重なっているかどうかで観測される強さが異なるため，容易ではない．歌声やドラム音も鳴っているときには，非高調波成分がどの音に起因するかの推定も加わり，より一層難しくなる．個々の音がいつ鳴り始めたか (発音時刻) を正確に推定したり音の数 (音源数) を推定したりすることも難しい課題である．

音楽ではこうした混合音理解に加え，複数の音の関係，時間構造を適切に理解する難しさもある．各楽器パートの前後の音との関係だけでなく，他の楽器パートとの関係，フレーズ単位の局所的な構造，楽曲全体の大局的な構造も重要である．しかし，こうした時間変化する現象や関係を適切に表現する技術は，未成熟である．さらに，さまざまな客観的な理解を，主観的な理解に結び付けて適切に扱うことも難しく，その方法は確立していない．

2.3 音楽の自動理解はどこまで可能になったか

音楽の自動理解技術は，さまざまな研究によって進歩し，従来の少数の音を人工的に合成した音響信号だけでなく，音楽配信や音楽 CD による楽曲のような，より実用性の高い実世界の複雑な音響信号も扱えるようになってきた．しかし，万能な音楽理解手法はまだ実現されておらず，あらゆる手法が，対象とする音楽 (入力音響信号) の属性に関して，必ず何らかの仮定を設けている．そして，その仮定によって問題の難易度は大きく変わることがあり，音楽の自動理解はどこまで可能になったかを，一概に述べることは難しい．例えば，音響信号がモノラルかステレオか，単一音か混合音か，混合音の場合には楽器数や同時発音数がいくつか，どのような音楽ジャンルか，リズムや楽曲構造がどれぐらい複雑か等のさまざまな属性が影響する．

現在流通している音楽では，ステレオの音響信号が主流である．しかし，ステレオ信号を対象として定位情報 (音が左右のどの方向から鳴っているか) を利用する研究事例 [9] [103] [167] [178] [192] は少なく，多くの研究では，その左右を平均したモノラルの音響信号を対象としている．これは，ステレオ信号に依存した手法がモノラル信号に適用できない一方，逆は適用でき，また人間もモノラル信号から音楽を理解できるため，音楽の自動理解を実現する上で本質的と考えられるからである．

人間が音楽を聴いたときのさまざまな主観的な理解は，その多様性や評価が難しいこともあってあまり研究されていない．一部，楽曲と感性語 (楽曲の印象を表す形容詞等) との対応付けは研究されている [159] [164]．一方，客観的な理解に関してはさまざまな自動理解の技術が提案されており，以下で紹介する．

2.3.1 一曲中で刻一刻と変化する事象の自動理解

音楽の自動理解の一つのイメージは，混合音を構成する個々の音を分離・抽出する音源分離や，個々の音符の高さや長さを推定する多重 F0 推定，小節中の位置まで推定して楽譜を作成する自動採譜である．それらの具体的事例は，文献 [89] に数多く紹介されている．楽譜情報に加えて，個々の音の音量や発音時刻等のより詳細な情報も含む標準 MIDI ファイル (SMF) を出力する研究も多い．例えば，対象とできる音楽は比較的単純なものに限られるが，音楽中のすべての音の発音時刻，音高，音量，音色等を同時に推定する

†3 「重なったまま理解する」とは，混合音を構成する個々の音の分離信号を得ずに理解することを意味し，例えば，2.3.1 節 で後述する「音楽情景記述」[58] [60] [61] では，そうした理解を目指している．

試みもある [16] [17] [31] [83] [84]．文献 [83] の手法では，個々の音の音量の時間変化までモデル化し，その組合せとして入力を表現する．高調波構造の仮定すらせずに推定する手法 [1] [143] [152] もある一方で，事前知識として SMF と組み合わせて，より複雑な音響信号を扱う手法 [71] [192] も存在する．文献 [192] の手法では，SMF によって個々の音符の高さは与えられるが，その発音時刻と音量・音色変化 (スペクトログラム) を詳細に求めることができる．

一方，「人間は，構成音の分離信号や楽譜に基づいて音楽を理解しているわけではない」という立場から，音楽的に訓練されていない「しろうと」の音楽理解の実現を目指す「音楽情景記述」(music scene description) [58] [60] [61] も提案されている．そこでは，構成音の分離信号や個々の音符を得ることに拘らずに，直接，メロディー，ベース，ビート，サビ等の音楽演奏中の情景を分析・理解した結果を記述する点が特徴的である．実は，仮に分離信号や個々の音符がすべて得られたとしても，こうした「しろうと」がわかる音楽演奏中の情景が得られるわけではないことに，注意する必要がある．

複数の楽器音が混在した混合音中のメロディーの音高推定は，最初に実現された PreFEst [58] [61] の他，複数の手法 [15] [36] [38] [39] [101] [116] [131] [139] が提案されている．基本的には，単位時間ごと (例えば 10ms ごとのフレーム単位) に音高を推定する手法が多いが，音符の遷移等を考慮して音符レベルで推定する手法 [38][39][116] も取り組まれている．ポピュラー音楽のようにメロディーが歌声の場合，歌声を事前にモデル化することで，歌声に特化して性能を向上することもできる [50] [139]．混合音中の歌声の基本周波数と各高調波の音量を推定すれば，その歌声単独 (主に母音) を再合成することもでき，例えば，歌詞のテキストが既知という条件の下で，歌詞と歌声とを時間的に対応付ける研究もある [21] [47] [48] [75]．しかし，歌詞が未知の歌声に対して音声認識のように歌詞を認識する技術は，まだ限定された条件下でしか実現されていない [191]．他には，ベースの音高推定に関しても，文献 [58] [61] [70] [140] のような手法が提案されている．

サビは楽曲を代表する重要な区間であり，それを推定する手法は数多く研究されてきた．その多くは，楽曲中で繰り返される類似区間 (フレーズ等) の検出に基づいている．例えば，楽曲の代表的な部分を一箇所切り出す手法 [11] [28] [98] や，主要な部分を残して短くする音楽要約手法 [30] [126]，すべてのサビ区間を網羅的に検出する手法 [63] 等が提案されている．この中で，最後に挙げたサビ区間検出手法 RefraiD [63] は，ポピュラー音楽を前提に転調の有無に関わらず各サビ区間の開始点，終了点を求めることができる特長を持っている．特徴量として何を用いるかによって，どのような観点から繰り返しが求まるかが決まる．例えば，代表的なクロマベクトル [63][†4] の場合には，コード進行が繰り返す区間が検出できることが多く，イントロとエンディングの繰り返しや，楽曲の一番，二番の繰り返しのような楽曲構造もある程度求まる [63] [111]．また，楽曲構造の区間に対応するラベルを推定する手法 [125] [135] や楽曲構造の区間境界の推定に特化した手法 [165] も提案されている．

ビートやテンポも音楽を聴く上で基本的な概念であり，人間が音楽に合わせて手拍子を打つように，各ビート (拍) を推定するビートトラッキング手法や，テンポを推定する手法が数多く提案されている [32] [59] [65] [69] [90] [142] [147] [170] [200]．階層的なビート構造を得ることで，小節レベルまで推定できる手法もある [59] [90] [170] [200]．ポピュラー音楽では比較的テンポ変化は少なく，曲全体で一つのテンポの値を求めることもあるが，クラシック音楽のようなテンポ変化が重要な音楽では，それに追従してビート・テンポを推定する必要がある [142]．問題設定は異なるが，楽譜が既知という条件の下で，楽譜と入力音響信号とを時間的に対応付けるスコアフォロイング (あるいはオーディオアラインメント) も実現されている [34] [74]．

他にも，混合音中の個々の音の楽器名を求める音源同定手法 [37][85][87][95] や，混合音中の個々の音に対してそうした楽器名を一意に特定せずに候補となるそれ

†4 スペクトルをオクターブごとに分割してそれらを足し合わせた表現で，例えば 12 次元のクロマベクトルは，12 音名の各音名の周波数におけるパワーを複数のオクターブに渡って加算して求める．

ぞれの楽器音であった可能性 (楽器音の存在確率) だけを求める手法 [88]，ドラム音の発音時刻を推定する手法 [55][59][65][110][121][179][182][200] も実現されている．楽曲の調を推定する手法 [19][56][77][78][93][124][169] コード進行を推定する手法 [14][92][93][148][180] もある．

2.3.2 一曲全体に対する自動理解

一曲全体の音響信号に対して，それがどんな音楽ジャンルであるかを自動的に推定する手法 [33][42][43][96][105][133][163][166][175] が提案されている．しかし，音楽ジャンルの分類には確立した定義があるわけではなく，研究によっても異なる．各楽曲の音響信号が持つ雰囲気や曲調を特徴量化し，事前に学習したジャンルごとの特徴量の分布に基づいて，その曲に一番近いジャンル名を求める手法が多い．そのためには，楽曲間の類似度あるいは距離を何らかの方法で計算することが基本となる．

楽曲間の類似度は，音楽情報検索でも重要であり，楽曲中の音色 (スペクトル包絡等) [5][43][166]，リズム [33][43][44][120][166]，歌声 [46][163]，構成音 [130] 等のさまざまな特徴に基づく楽曲間の類似度が提案されている．ハミング検索のように，メロディーの歌唱やハミングを検索キーとして，そのメロディーを持つ楽曲を検索する場合 [112] [153] [158] には，検索キーと混合音中のメロディーとの類似度を求める必要がある．楽曲間の類似度の特殊な例としては，ある曲が別の曲のカバー曲 (既存の楽曲を異なるアレンジで演奏した曲) であるかどうかを判断する手法 [13][40][57] もある．

混合音中の歌唱の理解に関しては，上記で述べた歌詞との時間的対応付け以外にも，歌手名を同定する手法 [10] [49] [161] [107] や，歌詞の言語を推定する手法 [146] [162] が提案されている．しかし，歌手の人数や歌い方の推定に関しては，まだほとんど研究がなされていない．伴奏なしの単独歌唱であれば，歌唱力の自動評価についても研究されている [220]．

専門家のより詳細な理解として，楽曲の音響信号からの演奏家や指揮者の同定，周波数特性やミキシング時の各種条件の推定が挙げられるが，まだほとんど研究されておらず，今後の課題である．

2.3.3 楽曲の集合に対する自動理解

アーティスト間の類似度 (距離) を，各アーティストの楽曲群間の類似度 (距離) に基づいて計算する取り組みもある [119]．そして，楽曲間の類似度やアーティスト間の類似度を応用し，特徴量をグルーピング (分類) することで，自動的にジャンルやスタイルに相当する代表的な楽曲集合・アーティスト集合を求めることができる [127]．

しかし，プレイリストの自動生成 [3][6][97][114][118] [122][123] に関する研究は多いものの，2.1.3 節 で述べたような，既存のアルバムやプレイリストの意図や曲の流れの自動理解は，まだ研究事例はほとんどない．アーティストとジャンルの詳細な関係に関する理解も，今後の課題である．

3 音声の認識・理解

音声は，我々人間にとって最も自然なコミュニケーション手段であり，音声認識・理解技術は，人間とコンピュータとの自然なインタラクションを実現するための基礎技術として，古くから研究されてきた．昔から SF 映画や小説等で度々取り上げられる人間と自由に会話するロボットやコンピュータのように，音声認識・理解研究における究極のゴールは，人間が音声を認識・理解する際の高度なメカニズムをコンピュータシステム上で実現することである．

近年，研究の進展とコンピュータの大幅な性能向上に伴い，自動音声認識・理解技術は格段の進歩を遂げ，その結果パソコン上のソフトウェアやカーナビ，携帯電話などで実用化もなされている．しかし，これらは主に大量のデータが処理可能になったことでもたらされたものであり，本質的な音声の認識・理解能力は人間のそれにはまだ遠く及ばないといえる．

3.1 音声に含まれる情報とは

前述したように，音声は人間にとって最も自然なコミュニケーション手段であるが，それゆえに音声という「メディア」を通じていかに日常的なコミュニケーションを成り立たせているか，音声という「音響信号」からどのような情報を得ているのか，ということを我々がふだん意識することは少ない．ここで

は，音声に含まれる情報を,「言語情報」と「非言語情報」[†5]の大きく 2 種類に分類することで議論を進める．

音声に含まれる言語情報とは，単語や文のような発話内容を直接的に表す情報のことを指し，人間同士の対話は基本的にこのような言語情報の交換によって成立する [206]．一方，非言語情報とは，言い淀みや韻律，怒りや喜びなどの発話者の心的状態など文字では書き起こせない情報を表し [206]，対話において効率性や自然性を持たせる働きを持っている．

音声研究においては,「認識」と「理解」の意味合い，目的などの違いは区別されているケースが多い (ただし厳密な使い分けは研究者によって見解が分かれる)．ここでの「認識」，すなわち「音声認識」とは，主として上記の言語情報を音声信号中から抽出することを意味し，音声信号というアナログの情報をテキスト (文字) というデジタルの情報に変換することにあたる．一方,「音声理解」とは,「認識」の上位概念として位置付けられ，単に音声信号を文字情報に変換するだけでなく，上記の非言語情報も考慮することで，発話者の意図や一連の会話の流れの理解，話題語 (キーワード) の抽出や発話内容の要約などを行うことを指す．

3.2 なぜ音声認識・理解は難しいか

我々がふだん当たり前のように行っている音声の認識・理解を，コンピュータシステム上で実現することは容易ではない．なぜなら，音声がさまざまな「ゆらぎ」を含む曖昧で多様な音響信号だからである．しかも，音声は音響信号としての曖昧性を持つうえに，言葉の意味を適切に理解するという，自然言語的な難しさもあわせ持つ．以下では，音声認識・理解を行う際の技術的な課題について整理していく．

3.2.1 音響的変動の影響

音声は，話者それぞれ発声方法や発声器官が少しずつ異なるために，たとえ同じ単語や音素であっても話者によってその音響的な性質が異なる．また，同じ話者の音声でもそのときの感情や発声した環境によってその音響的性質は変動する．

実環境においては，対象となる音声のみが入力されるような「クリーン」な収録条件であることはほとんどなく，車の中，家庭内 (キッチン等)，駅や空港など，音声信号以外の雑音の影響が著しい．この場合，音声信号中に雑音信号が重畳することで混合音となり，大きな音響的変動が発生する．このようにシステムに入力される信号が音声だけとは限らないので，発声された音声の区間を正しく検出する処理も必要となる．

3.2.2 言語的制約の必要性

音響的特徴のみによる照合だけでは，同音異義語あるいは音響的に類似した単語を識別し，言語的に意味のある認識結果を得ることが難しい．そこで，何かしら言語的な制約を与えることによって，文法や構文からの逸脱を回避し，音響的な曖昧性を解消する必要がある．

音声認識システムにおいて一般的に用いられている言語的制約とは，単語と単語の接続関係 (確率) を表したものであり，通常，大量のテキストデータから学習される．この言語的制約が，想定する発話内容を反映したものでなければ，正しい認識結果は得られない．また，言語的制約内で定義されていない単語は，システムの知らない単語 (未知語) として扱われ，基本的に認識はできない．

3.2.3 発声方法の違い

一口に音声認識といっても，発声方法の違いによってその形態，難しさは大きく異なる．最も認識しやすい発声方法としては，1 回の入力で 1 単語のみを発声する方式である．この場合，システムに登録された単語のうち，音響的に最も近い単語を認識結果として出力する問題となる．また,「私/は/今日/大学/に/行く」のように，文を単語ごとに区切って (単語と単語の間に一定の間を空ける) 発声する方式もある．しかし，人間の日常会話においてはこれらの発声方法は不自然なものであり，単語の境界を特に意識せずに，連続して発声された音声を認識する「連続音声認識」の形態が望ましい．ただし，連続発声された音声信号には単語や音素といった明確な区切りがないため，信号中のどこからどこまでが単語，あるいは音素なのか

†5 ここでは便宜上，パラ言語情報 [215] は非言語情報に含まれるものとする．

を自動的に区分化する処理が必要となる．

3.2.4 自然な発話への対処

音声は発話スタイルにより,「読み上げ音声」と「自然発話音声」の大きく二つに分類できる．新聞記事などのあらかじめ用意されたテキストの読み上げ音声であれば高い認識率を得られる認識システムでも，人間同士の日常会話などの自然発話音声をそのまま認識しようとすると大幅な性能劣化が生じる．しかし，人間同士のコミュニケーションにおいて，読み上げ音声は極めて不自然である．人間とコンピュータの間の自然な音声インタフェースを実現するには，自然発話音声を頑健に認識する技術が重要となる．

このような自然発話音声においては，間投詞，言い淀み，言い直し，文法を逸脱した文，省略等による不完全な文が発声されるケースが多い．これらの現象には，読み上げ音声に対する言語的制約では対応できない．また，発音の変形などが頻繁に発生し，音響的にもゆらぎが大きく難しくなる．

3.2.5 「理解」することの難しさ

人間同士の会話，コミュニケーションにおいて最も重要なのは，お互いに相手の意図を理解し合うことである．コンピュータでも音声信号をテキストに変換するだけでなく，発声者の意図を音声信号中から抽出する「音声理解」が重要となる．その実現のためには，これまでに述べた音声信号をテキストに変換する「音声認識」における問題点に対処するとともに，発声者の意図に相当する情報を何らかの方法で同定・抽出しなければならない．

3.3 音声の自動認識・理解はどこまで可能になったか

音声の自動認識技術は，国内外の多くの研究によって格段に進歩し，現在もさまざまな研究がなされている．とくに近年は，放送ニュース，会議，講演などの実世界・実環境における音声データを対象とした，大語彙連続音声認識 (Large Vocabulary Continuous Speech Recognition: LVCSR) に関する研究が中心的テーマとなっている．大語彙連続音声認識とは，数千語以上の語彙 (単語) を扱う連続音声認識を指し，ここではこれを単に音声認識と呼ぶことにする．

以下では，現在の一般的な音声認識システムの仕組みについて概観し，各要素技術における最近の研究事例を紹介する．

3.3.1 音声認識システムの概観

図 1 に典型的な音声認識システムの構成図を示す．現在の音声認識システムは，基本的に，確率モデルを用いた統計的手法に基づいており，言語によらず世界共通でほぼ同一の枠組みが採用されているといってよい．

音声分析は，認識処理の前段階として，マイクロフォンから入力された音声信号から音声認識に有用な情報 (主としてスペクトル包絡) を特徴ベクトルとして抽出する処理である．

音声認識は，与えられた特徴ベクトル O に対して，任意の単語列 W の中から，最も適合する (事後確率 $P(W|O)$ が最大となる) 単語列 $\hat{W}$ を求める処理として以下のように定式化される．

$$\begin{aligned}\hat{W} &= \operatorname*{argmax}_W P(W|O) \\ &= \operatorname*{argmax}_W \frac{P(O|W)P(W)}{P(O)} \\ &= \operatorname*{argmax}_W P(O|W)P(W)\end{aligned}$$

ここで，W に対する最大化において，$P(O)$ は無視できる．

$P(O|W)$ は，単語列 W に対する音響パターンが O である確率を表し，音響モデルと呼ばれる確率モデルにより計算される．音響モデルは，あらゆる文，単語パターンの音響確率を計算できるように，ほとんどの場合は，音素に代表される音韻単位で用意される．現在では，特に大語彙，不特定話者，連続音声などの条件において音響的な変動をうまく表現可能なモデルとして,「隠れマルコフモデル (Hidden Markov Model: HMM)」が広く利用されるようになった [91]．HMM は，定常な信号源を確率的に切り替えていくことにより，音声のような時々刻々と変化していく非定常信号を表現できる特長を持つ．

言語モデルは，音声認識を行う際の言語的制約を規定するものである．従来，言語モデルとしては，専門家が人手で記述した単純な文法を利用することが主流であったが，大語彙の音声認識においては人手での

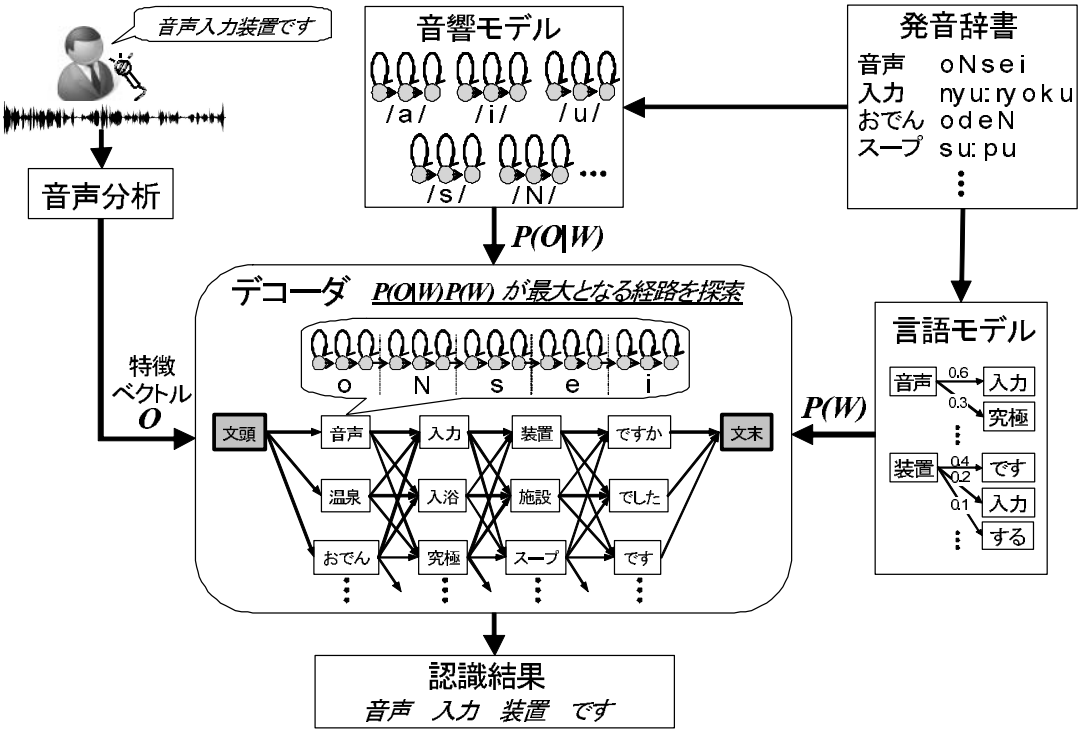

図 1　典型的な音声認識システムの構成

文法記述は現実的ではない．そこで，電子化された大量の言語データ (テキストコーパス) からモデルを統計的に推定するアプローチが主流となっている．現在の音声認識システムに最もよく用いられている言語モデルは N-gram と呼ばれる単純な確率モデルである．N-gram は，ある時点での単語の生起確率は直前の $N-1$ 個の単語のみに依存すると仮定し，N 個の単語のつながりを確率化したものとして定義される．ただし，日本語の場合，欧米語と違って単語の定義が明確ではないため，事前に何らかの形でコーパス中の文を単語単位に分割する必要がある．単語分割には，形態素解析器 [209] [219] が利用されることが多く，ほとんどの日本語音声認識システムでは，単語の単位として形態素が採用されている．

発音辞書は，システムに登録されている語彙 (言語モデル内で規定された語彙) に対する読み (発音) 情報を与えるものである．発音辞書は，音響モデルと言語モデルの間の掛け橋となるものであり，各単語の読み情報に従って音韻単位の音響モデルが連結され，単語単位の音響モデルが構成される．

デコーダ (認識エンジン) は，入力音声 O に対して音響的にも言語的にも最もマッチする単語列 W(すなわち $P(O|W)P(W)$ が最大となる単語列 W) を，膨大な単語列の組み合わせの中から探索する処理のことを指す．したがって，デコーダにどのような探索アルゴリズムを採用するかは，システム全体の処理時間に大きく影響する．とくに大語彙連続音声認識の場合は探索空間が膨大であるため，探索アルゴリズムをいかに設計するかは重要な問題となる．

3.3.2　発話区間検出，音響セグメンテーション

放送ニュース [22][27][177]，会議 [79][108][181]，ポッドキャスト [66][113][196][197][204] など，実環境のデータにおいては，音声だけでなく，音楽，雑音などさまざまな「音響イベント」が入力音響信号中に存在する．このようなデータに対して音声認識を行うためには，認識前にあらかじめ，音声の存在する区間を検出しておく必要がある．そのために，音響信号を各音響イベントごとに区分けする音響セグメンテーション

が検討されている [18] [23] [53] [68]．音響セグメンテーション手法としては，想定する各音響イベントごとにその音響的特徴を表す確率モデルを学習しておき，入力された音響信号に対して最尤のイベントのモデルを識別結果とするモデルベース [53] [68]，音響的な特徴量をもとに入力音響信号中の局所的な変化を算出することで，イベントの境界点をボトムアップに求めていくメトリックベース [18] [23] の 2 種類に分類される．モデルベースのアプローチは高い精度を示す反面，事前学習に基づくため，事前に想定した音響イベントしか検出できない．一方，メトリックベースは，精度はモデルベースより低いものの，事前に音響イベント数を限定しないアプローチなので，未知のイベントにも対応可能である．

3.3.3 音響モデル

音響モデルの高度化に関する研究は，近年も精力的に行われている．与えられた学習データに対して最適なモデル・構造を自動選択・最適化する，いわゆるモデル選択問題に関してはこれまでにさまざまな提案がなされ [149] [172]，音響モデルの頑健性の向上に大きく貢献している．音響モデルの単位としては，音素単位が一般的であるが [91] [145] [210]，日本語は英語などに比べて音節数が極端に少なく，音響モデルの単位として音節を用いることの有効性が示されている [195] [208]．

また，実環境の音声を対象にした場合，音響モデルの学習用データと実際の認識時のデータとの間に，さまざまな音響的なミスマッチ (例えば話者の違い，雑音の影響など) が発生する．したがって，実環境音声認識システムにおいては，時々刻々と変化する音響変動に「適応」していく技術が重要となり，これまでにもさまざまな適応手法が提案されている [52] [94]．

3.3.4 言語モデル

言語モデル (N-gram) の高度化に関する研究としては，3 単語連鎖 ($N = 3$) 程度の局所的な情報だけでなく，より長いスパンの関係を考慮するトリガーモデル [137] [67] や，潜在的意味解析による言語モデル [12] などが提案されている．また，従来の N-gram モデルであっても，言語モデル学習用のテキストデータ量の増大が認識性能向上に大きくつながることが報告されている [157]．

放送ニュースやポッドキャストなどの，日々話題が移り変わり，新たな話題が増え続けるような実世界の音声データに対しては，言語モデルを日々更新することで，最新の話題を常にカバーできる状態に保つ必要がある [197] [198]．例えば，ポッドキャストの音声認識のために，日々更新される Web ニュースサイトのテキストを収集して学習に利用することで，増え続ける話題に頑健な認識手法 [197] が提案されている．

3.3.5 発音辞書

前述したように，発音辞書はシステムに登録された語彙の発音を与えるものであり，発音辞書中に定義されていない単語は認識処理で一切扱うことはできない (知らない単語，読みがわからない単語は認識できない)．このような発音は，小語彙の音声認識においては人手で記述することも可能であるが，大語彙の音声認識の場合は現実的ではない．そこで一般的に，日本語の大語彙連続音声認識の場合，形態素解析器が出力した (形態素解析辞書に定義されている) 読み情報をそのまま利用することが多かった．

ただし，ポッドキャストなどの実世界の音声データでは，内容は多岐にわたり，しかも世の中の最新の動向や話題について話されていることも多く，汎用的な形態素辞書ですべてをカバーすることは不可能である．そこで，Web 上のキーワード辞書サービスから，最新の単語に対する読み情報を自動的に獲得する方法が提案されている [197] [216]．

また，自然発話音声特有の現象である発音変動を，発音辞書において音韻レベルのシンボリックな系列の並びによってモデル化する研究もおこなわれている [150] [151] [193] [218]．

3.3.6 話者認識

音声の音響信号に含まれる情報の中から「話者性」のみを抽出する技術である「話者認識」は，個人認証，セキュリティシステムへの応用技術として重要視されるなど，音声認識技術と並行して古くから盛んに研究がなされてきた [134] [211]．話者認識技術は，入力された音声が本人であるかどうかを単に受理/棄却する「話者照合」，入力された音声がどの話者であるかを同定する「話者識別」に大きく分けることができ

る．いずれにおいても，HMM 等の確率モデルに基づくアプローチが現在では主流となっている．これは，各話者ごとにその音声の特徴を表す確率モデル (話者モデル) を HMM により事前に学習・生成しておき，入力された話者の音声と各話者モデルを照合することで，声の近さに基づく判定を実現するものである．確率モデルによるアプローチは，モデルの学習を行うために対象となる各話者の音声を事前に収集する必要があるが，十分な量の学習データ (各話者ごとに数十秒から数分) があれば，高い精度を達成することができる．

また最近では，このような話者認識と，3.3.2 節で述べた音響セグメンテーションの両者を包含する課題として,「話者決定 (speaker diarization)」が注目されている [160] [211]．これは，放送ニュースや会議など，音声以外の雑音等も含む比較的長時間の音響信号に対して,「どの話者がいつ発声したか」を自動決定するものである [211]．話者決定結果は，後段の音声認識処理において精度向上のために活用できるだけでなく，音声のアーカイブに対するインデキシングのためにも利用でき，さまざまな研究が進められている [41] [160]．

3.3.7 音声認識から音声理解へ

究極の音声認識・理解は，音声信号から言語情報を抽出しつつ，発声者の意図を正しく理解することである．音声中のどのような情報を意図とするのか，またどこまでの理解が得られれば目的が達成されたといえるのかを定義することは難しいが，これまでにも音声理解を目指した萌芽的な研究が行われている．

例えば，音声中もしくは音声認識結果のテキスト情報から，キーワードやキーフレーズ，話題を抽出する研究 [86][185][189] も，簡易的な音声理解といえる．また，キーワードよりもさらに言語的に豊かな意図の表現形式として，音声の「要約」があり，これを音声認識結果から自動的に生成する研究もなされている [72][213][217]．音声要約は，音声中の話題を担っている箇所 (単語あるいは文) をできるだけ文法的な制約に従う形で抽出しつつ，話題に関連のないそれ以外の箇所を削ぎ落としていく処理と位置づけられる．このような技術は，議事録作成，アーカイビングなどの用途においても有用である．

一方，音声コミュニケーションという意味では，上記のような発話内容を端的に表す情報以外にも，音声音響信号中に頻出する各種非言語情報をいかにして扱うかが重要となる．例えば，イントネーションやアクセントなどの韻律情報は，話し手の意図や態度の伝達に重要な役割を担うことから，これまでにもさまざまな分析がなされており [51] [187]，音声対話システムの対話制御への利用 [206] や韻律を考慮した音声認識手法 [188] に関する研究も行われている．また，音声音響信号から，喜び，笑い，悲しみなど感情を識別する研究も進められており [171]，声の高さやスペクトル等の特徴量を利用して，HMM をベースとした識別器を学習する手法等が報告されている [144] [176]．感情識別は，音声対話や音声理解における重要な要素技術の一つであるだけでなく，ゲーム等のエンターテイメントへの応用も多くなされている．

従来の音声技術は，基本的にはユーザとコンピュータ間の 1 対 1 でのインタラクションを想定したものが多い．しかし，実環境での利用場面では，複数のユーザがコンピュータとインタラクションを行うという状況も考えられる．そのような状況において，円滑な音声対話，音声理解を実現するためには，ユーザの発声がコンピュータに向けられたものなのか，ユーザ間の対話なのかを区別できる技術が必要となる．文献 [194] では，音声認識結果として求められる言語的な特徴 (発話固有の言い回し，単語) を利用した識別器を用いることで，ロボットへの発話とユーザ間の雑談を高い性能で区別できている．一方,「音声スポッタ」 [203] というインタフェースでは，コンピュータに向けて発声する際に,「母音を延ばして言い淀んだ後に故意に高い声で発声」する特殊で不自然な言い方をルール化することで，発話内容によらず，信号処理 (言い淀み検出，音高推定) のみで区別することを可能にしている．

現状では，前節までで述べたように，実環境音声に対していかに頑健に認識を行うか，についての課題が多く残されているため，これらが主として音声研究における中心的なテーマとなっている．しかし，今後，真の意味で人間とコンピュータ間のコミュニケー

ションを実現するためには，上記のような音声理解を指向した技術の重要性がますます高まってくると考えられる．

4 研究で活用できるツールの紹介

音楽・音声の音響信号を認識・理解する研究の成果として，ツール (ソフトウェア) の形で直接提供されているものは多くはないが，本章では，インタフェースやインタラクションの研究で活用できそうな代表的なツールを紹介する．同じツールが音楽と音声の両方に使える場合も多いが，以下では主に音響信号全般を扱う際のツールを紹介した後に，音楽音響信号向けツール，音声音響信号向けツールを紹介する．最後に，統計的学習や評価等で重要となるデータベースについても紹介する．

4.1 音響信号全般を扱うためのツール

音響信号を記録したファイルを編集したり可視化したりするツールとして，

- WaveSurfer [173]
- Audacity [7]
- Ardour [4]

等が挙げられる．統計的手法で隠れマルコフモデル (HMM) を使用する際には，そのツールキット

- HTK [73]

が使われることが多い (特に音声の研究では頻繁に使用される)．機械学習には，データマイニングソフトウェア

- Weka [174]

が便利である．商用システムではあるが，科学技術計算で用いられることの多い数値解析ソフトウェア

- MATLAB [104]

も，音響信号処理の研究ではよく用いられる．

4.2 音楽音響信号を扱うためのツール

音楽情報処理で便利な各種音響特徴量を抽出できるツールとして，

- MARSYAS [102]
- CLAM [25]
- Sonic Visualiser [154]
- jMIR [80]
- Vamp audio analysis plugin system [168]

等が挙げられる．これらは，音響分析やアノテーションも可能な統合環境として提供されている場合が多い．リアルタイムな音響分析・合成では，

- ChucK [24]
- Audicle [8]

のような優れた環境も存在する．MATLAB 上では，音楽用音響特徴量の抽出が可能な

- MIRtoolbox [109]

が便利である．他にも，音楽情報検索技術の評価を主目的とした

- M2K [100]

や，音楽情報処理研究のための XML ベースの共通データフォーマット

- CrestMuseXML [29]

用の音響信号処理用 Toolkit が開発中である．

4.3 音声音響信号を扱うためのツール

音声情報処理においては，主に研究者や開発者を対象としたツール，ソフトウェアの整備が精力的に進められている．まず，3.3.1 節でも述べた，大語彙も扱うことが可能な音声認識システム (認識エンジン) としては，

- HTK [73]
- Julius [81]
- Sphinx [155]
- RWTH ASR [138]

等が挙げられる．音声認識用の音響モデル (HMM) の構築・学習に関しては，HTK が完成度が高く，世界中の研究機関で利用されている．Sphinx，RWTH ASR においても，音響モデルの構築は可能である．

統計的言語モデル (N-gram) の構築には，

- SRILM [156]
- CMU-Cambridge Statistical LM Toolkit [26]
- Palmkit [117]

等が利用可能である．N-gram に関しては，共通のフォーマット「ARPA 形式」が利用されることが多く，上記の三つのツールも ARPA 形式を扱うことが可能である．また，日本語の言語モデルを作成する場

合には，単語分割のために形態素解析器が不可欠である．代表的なものとしては，

- Mecab [106]
- ChaSen [20]
- Juman [82]

等が挙げられる．

4.4 データベース，コーパス

音楽では，音声と比べてデータベースの整備が遅れているが，研究目的での自由な利用のために構築された世界で最初の大規模な音楽データベース

- RWC 研究用音楽データベース [205]

が頻繁に使われている．多様な音楽ジャンルの楽曲 315 曲と約 50 種 150 個体の楽器の演奏音が収録されており，音響信号と標準 MIDI ファイル，歌詞のテキストファイルが用意されている．また，これらの楽曲に対するメロディーやビート等のアノテーション (正解ラベル)

- AIST アノテーション [62]

も公開されている．楽器音データベースに関しては，

- McGill University Master Samples [115]
- Iowa Musical Instrument Samples [45]

等もよく用いられる．ドラム演奏に特化したデータベース

- ENST-Drums [54]

や，同一歌詞を 72 名が歌唱した試料集

- 日本語を歌・唄・謡う [183]

も公開されている．他にも，音楽情報検索技術の評価の枠組み MIREX [35] では，個々の評価タスクに応じてデータセットが配布されることがある．

一方，音声研究では，古くからデータベース，コーパスの利用が重要視され，それらの整備が精力的に進められてきた．ここでは日本語の音声言語データベース，コーパスについて紹介する．大語彙連続音声認識の研究を目的とした，大規模な読み上げ音声コーパスとしては，

- 日本音響学会読み上げ音声コーパス (ASJ-JNAS) [76]

がある．また，ATR では，読み上げや自然発話など，さまざまな種類の音声コーパスが研究成果物としてリリースされている (例えば [212])．最近では国内でも，より実環境に近い音声を収録したコーパスの整備が重要視され，

- 日本語話し言葉コーパス (CSJ) [186] [214]
- CIAIR 実走行車内音声データベース [199]

等が構築された．前者は学会講演音声を収録したもの，後者は実走行車内での対話音声を収録したものである．

5 おわりに

本解説論文では，音楽の音響信号，音声の音響信号を認識・理解する研究の現状を紹介し，さまざまな研究事例を挙げてきた．これらの研究は大きく進展してきたものの，そこで実現された認識・理解技術をインタフェースやインタラクション研究に用いる場合には，その性能が 100%ではないことに注意する必要がある．多くの技術では，70〜90%程度のことが多く，100%に近づけようとすればするほど，技術的な難易度は非常に高くなる．これは，人間も聞き間違いをするように音響信号を扱う上での根本的な難しさがあると共に，技術的にまだ未熟なところがあることも要因となっている．そこで，そうした技術が誤りを起こすことを前提とした活用方法を考えることが重要となる．

今後の研究の方向性として，音楽と音声を区別せずに，それらの統合的理解を目指す研究も増えてくることを期待したい．音楽か音声かを識別する研究事例 [2] [128] [129] [141] は多く，音声の背景に音楽が鳴っている場合等の難しい課題に取り組んでいる研究 [99] や同じ人間の声でも歌声と話し声を識別する研究 [190] もあるが，両者の統合的な理解はまだまだこれからである．実際には，同じ音を扱うということで共通した技術が用いられる場合もあるが，音楽理解では入力を混合音として扱うことが多く，音声理解では単一音あるいはそれに雑音が加わった音として扱うことが多い．そのために，両者を同じ枠組みで扱う研究は少なかった．しかし今後は，音楽と音声の一方の分野で培われた技術を他方の分野に応用したり，両方の分野の技術を結集しなければ解決できない課題に取り組んだりする試みが増えていくと考えられる．例えば，

音楽の混合音中の歌声が何を歌っているか，という歌詞の自動認識は，音楽理解問題として難しいだけでなく，音声の観点からも最も難しいクラスの音声認識問題として位置付けられ，その解決には両方の分野の技術が不可欠となる．

本解説論文ではすべてのトピックは網羅できなかったが，他の解説や書籍として，音楽では文献[64][89][136][201][202] 等が，音声では文献[132][184][207] 等が参考となる．また，こうした分野の最新の研究成果を知るためには，学会の出版物をサーベイすることが不可欠である．学会論文誌では，

- 情報処理学会論文誌
- 電子情報通信学会論文誌
- 日本音響学会誌
- 音楽知覚認知研究
- Speech Communication
- Journal of New Music Research
- Computer Music Journal
- Computer Speech and Language
- IEEE Transactions on Audio, Speech, and Language Processing
- EURASIP Journal on Advances in Signal Processing

等に関連成果が掲載されることが多い．代表的な国際会議やワークショップには，

- ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing
- ISMIR: International Conference on Music Information Retrieval (2000-2001 年は International Symposium on Music Information Retrieval)
- ICMC: International Computer Music Conference
- ICMPC: International Conference on Music Perception and Cognition
- Interspeech: 偶数年開催の ICSLP (International Conference on Spoken Language Processing) と奇数年開催の Eurospeech (European Conference on Speech Communication and Technology) が 1999 年から統合
- WASPAA: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics
- SAPA: ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (2004 年は最後が Perceptual Audio Processing)

等が挙げられる．国内の最新の研究活動は，研究会

- 情報処理学会 音楽情報科学研究会 (SIGMUS)
- 日本音響学会 音楽音響研究会 (MA)
- 情報処理学会 音声言語情報処理研究会 (SIGSLP)
- 電子情報通信学会／日本音響学会 音声研究会 (SP)
- 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD)

や，春秋の研究発表会・全国大会

- 日本音響学会 研究発表会
- 日本音楽知覚認知学会 研究発表会
- 情報処理学会 全国大会
- 電子情報通信学会 総合大会
- 情報科学技術フォーラム (FIT)

等で紹介されることが多い．今後も，より多くの研究者が取り組むことで音楽・音声の認識・理解研究が発展し，同時に，その成果を活用したインタフェース・インタラクション研究もより一層盛んになることを期待したい．

参考文献

[1] Abdallah, S. A. and Plumbley, M. D.: Polyphonic Music Transcription by Non-Negative Sparse Coding of Power Spectra, in *Proc. of ISMIR 2004*, 2004.

[2] Ajmera, J., McCowan, I. and Bourlard, H.: Robust HMM-Based Speech/Music Segmentation, in *Proc. of ICASSP 2002*, Vol. 1, 2002, pp. 297–300.

[3] Alghoniemy, M. and Tewfik, A. H.: A Network Flow Model for Playlist Generation, in *Proc. of ICME 2001*, 2001.

[4] Ardour: http://www.ardour.org/.

[5] Aucouturier, J.-J. and Pachet, F.: Music Similarity Measures: What's the Use?, in *Proc. of ISMIR 2002*, 2002, pp. 157–163.

[6] Aucouturier, J.-J. and Pachet, F.: Scaling Up Music Playlist Generation, in *Proc. of ICME 2002*, 2002.

[7] Audacity: http://audacity.sourceforge.net/.

[8] Audicle: http://audicle.cs.princeton.edu/.

[9] Barry, D., Lawlor, B. and Coyle, E.: Sound

Source Separation: Azimuth Discrimination and Resynthesis, in *Proc. of DAFX-04*, 2004, pp. 240–244.

[10] Bartsch, M. A.: *Automatic Singer Identification in Polyphonic Music*, PhD Thesis, The University of Michigan, 2004.

[11] Bartsch, M. A. and Wakefield, G. H.: To Catch A Chorus: Using Chroma-based Representations for Audio Thumbnailing, in *Proc. of WASPAA'01*, 2001, pp. 15–18.

[12] Bellegarda, J. R.: Large Vocabulary Speech Recognition with Multispan Statistical Language Models, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 8, No. 1(2000), pp. 76–84.

[13] Bello, J. P.: Audio-Based Cover Song Retrieval Using Approximate Chord Sequences: Testing Shifts, Gaps, Swaps and Beats, in *Proc. of ISMIR 2007*, 2007.

[14] Burgoyne, J. A., Pugin, L., Kereliuk, C. and Fujinaga, I.: A Cross-Validated Study of Modelling Strategies for Automatic Chord Recognition in Audio, in *Proc. of ISMIR 2007*, 2007.

[15] Cao, C., Li, M., Liu, J. and Yan, Y.: Singing Melody Extraction in Polyphonic Music by Harmonic Tracking, in *Proc. of ISMIR 2007*, 2007.

[16] Cemgil, A. T.: *Bayesian Music Transcription*, PhD Thesis, Radboud University of Nijmegen, 2004.

[17] Cemgil, A. T., Kappen, B. and Barber, D.: A Generative Model for Music Transcription, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 2(2006), pp. 679–694.

[18] Cettolo, M., Vescovi, M. and Rizzi, R.: Evaluation of BIC-based Algorithms for Audio Segmentation, *Computer Speech & Language*, Vol. 19, No. 2(2005), pp. 147–170.

[19] Chai, W. and Vercoe, B.: Detection of Key Change in Classical Piano Music, in *Proc. of ISMIR 2005*, 2005.

[20] ChaSen: http://chasen.naist.jp/hiki/ChaSen/.

[21] Chen, K., Gao, S., Zhu, Y. and Sun, Q.: Popular Song and Lyrics Synchronization and Its Application to Music Information Retrieval, in *Proc. of MMCN'06*, 2006.

[22] Chen, S. S., Eide, E. M., Gales, M. J. F., Gopinath, R. A., Kanevsky, D. and Olsen, P. A.: Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News, in *Proc. of ICASSP'99*, Vol. 1, 1999, pp. 37–40.

[23] Chen, S. S. and Gopalakrishnan, P. S.: Clustering via the Bayesian Information Criterion with Applications in Speech Recognition, in *Proc. of ICASSP'98*, Vol. 2, 1998, pp. 645–648.

[24] ChucK: http://chuck.cs.princeton.edu/.

[25] CLAM: http://clam.iua.upf.edu/.

[26] CMU-Cambridge Statistical LM Toolkit: http://mi.eng.cam.ac.uk/ prc14/toolkit.html.

[27] Cook, C. D., Kershaw, D. J., Christie, J. D. M., Seymour, C. W. and Waterhouse, S. R.: Transcription of Broadcast Television and Radio News: The 1996 Abbot System, in *Proc. of ICASSP'97*, Vol. 2, 1997, pp. 723–726.

[28] Cooper, M. and Foote, J.: Automatic Music Summarization via Similarity Analysis, in *Proc. of ISMIR 2002*, 2002, pp. 81–85.

[29] CrestMuseXML: http://www.crestmuse.jp/cmx/.

[30] Dannenberg, R. B. and Hu, N.: Pattern Discovery Techniques for Music Audio, in *Proc. of ISMIR 2002*, 2002, pp. 63–70.

[31] Davy, M. and Godsill, S. J.: Bayesian Harmonic Models for Musical Signal Analysis, in *Bayesian Statistics 7*, 2003, pp. 105–124.

[32] Dixon, S.: Automatic Extraction of Tempo and Beat from Expressive Performances, *Journal of New Music Research*, Vol. 30, No. 1(2001), pp. 39–58.

[33] Dixon, S., Pampalk, E. and Widmer, G.: Classification of Dance Music by Periodicity Patterns, in *Proc. of ISMIR 2003*, 2003, pp. 159–165.

[34] Dixon, S. and Widmer, G.: MATCH: A Music Alignment Tool Chest, in *Proc. of ISMIR 2005*, 2005.

[35] Downie, J. S.: The Music Information Retrieval Evaluation Exchange (2005–2007): A Window into Music Information Retrieval Research, *Acoustical Science and Technology*, Vol. 29(2008), pp. 247–255.

[36] Durrieu, J.-L., Richard, G. and David, B.: Singer Melody Extraction in Polyphonic Signals Using Source Separation Methods, in *Proc. of ICASSP 2008*, 2008, pp. 169–172.

[37] Eggink, J. and Brown, G. J.: A Missing Feature Approach to Instrument Recognition in Polyphonic Music, in *Proc. of ICASSP 2003*, 2003, pp. V–553–556.

[38] Eggink, J. and Brown, G. J.: Extracting Melody Lines from Complex Audio, in *Proc. of ISMIR 2004*, 2004, pp. 84–91.

[39] Ellis, D. P. W. and Poliner, G. E.: Classification-Based Melody Transcription, *Machine Learning Journal*, Vol. 65, No. 2–3(2006), pp. 439–456.

[40] Ellis, D. P. W. and Poliner, G. E.: Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking, in *Proc. of ICASSP 2007*, 2007, pp. IV–1429–1432.

[41] Fergani, B., Davy, M. and Houacine, A.: Speaker Diarization Using One-class Support Vector Machines, *Speech Communication*, Vol. 50, No. 5(2008), pp. 355–365.

[42] Flexer, A.: A Closer Look on Artist Filters for Musical Genre Classification, in *Proc. of ISMIR 2007*, 2007.

[43] Flexer, A., Gouyon, F., Dixon, S. and Widmer, G.: Probabilistic Combination of Features for Music Classification, in *Proc. of ISMIR 2006*, 2006.

[44] Foote, J., Cooper, M. and Nam, U.: Audio Retrieval by Rhythmic Similarity, in *Proc. of ISMIR 2002*, 2002, pp. 265–266.

[45] Fritts, L.: University of Iowa Musical Instrument Samples, http://theremin.music.uiowa.edu/MIS.html.

[46] Fujihara, H. and Goto, M.: A Music Information Retrieval System Based on Singing Voice Timbre, in *Proc. of ISMIR 2007*, 2007.

[47] Fujihara, H. and Goto, M.: Three Techniques for Improving Automatic Synchronization between Music and Lyrics: Fricative Sound Detection, Filler Model, and Novel Feature Vectors for Vocal Activity Detection, in *Proc. of ICASSP 2008*, 2008, pp. 69–72.

[48] Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T. and Okuno, H. G.: Automatic Synchronization between Lyrics and Music CD Recordings Based on Viterbi Alignment of Segregated Vocal Signals, in *Proc. of IEEE ISM 2006*, 2006, pp. 257–264.

[49] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection, in *Proc. of ISMIR 2005*, 2005, pp. 329–336.

[50] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search, in *Proc. of ICASSP 2006*, 2006, pp. V–253–256.

[51] Fujisaki, H.: Prosody, Models, and Spontaneous Speech, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, Springer, 1997, pp. 27–42.

[52] Gauvain, J. and Lee, C.: Maximum A-posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 2(1994), pp. 291–298.

[53] Gauvain, J.-L., Lamel, L. F. and Adda, G.: Partitioning and Transcription of Broadcast News Data, in *Proc. of ICSLP'98*, Vol. 4, 1998, pp. 5121–5124.

[54] Gillet, O. and Richard, G.: ENST-Drums: An Extensive Audio-Visual Database for Drum Signals Processing, in *Proc. of ISMIR 2006*, 2006.

[55] Gillet, O. and Richard, G.: Supervised and Unsupervised Sequence Modelling for Drum Transcription, in *Proc. of ISMIR 2007*, 2007.

[56] Gómez, E.: *Tonal Description of Music Audio Signals*, PhD Thesis, Pompeu Fabra University, 2006.

[57] Gómez, E. and Herrera, P.: The Song Remains the Same: Identifying Versions of the Same Piece Using Tonal Descriptors, in *Proc. of ISMIR 2006*, 2006.

[58] Goto, M.: A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals, in *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, 1999, pp. 31–40.

[59] Goto, M.: An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds, *Journal of New Music Research*, Vol. 30, No. 2(2001), pp. 159–171.

[60] Goto, M.: Music Scene Description Project: Toward Audio-based Real-time Music Understanding, in *Proc. of ISMIR 2003*, 2003, pp. 231–232.

[61] Goto, M.: A Real-time Music Scene Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, *Speech Communication*, Vol. 43, No. 4(2004), pp. 311–329.

[62] Goto, M.: AIST Annotation for the RWC Music Database, in *Proc. of ISMIR 2006*, 2006.

[63] Goto, M.: A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 5(2006), pp. 1783–1794.

[64] Goto, M. and Hirata, K.: Recent Studies on Music Information Processing, *Acoustical Science and Technology*, Vol. 25, No. 6(2004), pp. 419–425.

[65] Goto, M. and Muraoka, Y.: A Beat Tracking System for Acoustic Signals of Music, in *Proc. of ACM Multimedia'94*, 1994, pp. 365–372.

[66] Goto, M., Ogata, J. and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, in *Proc. of Interspeech 2007*, 2007.

[67] GuoDong, Z. and KimTeng, L.: Interpolation of N-gram and Mutual-information Based Trigger Pair Language Models for Mandarin Speech Recognition, *Computer Speech & Language*, Vol. 13, No. 2(1999), pp. 125–141.

[68] Hain, T., Johnson, S. E., Tuerk, A., Woodland, P. C. and Young, S. J.: Segment Generation and Clustering in the HTK Broadcast News Transcription System, in *Proc. of 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 133–137.

[69] Hainsworth, S. and Macleod, M.: Beat Tracking With Particle Filtering Algorithms, in *Proc. of WASPAA 2003*, 2003, pp. 91–94.

[70] Hainsworth, S. W. and Macleod, M. D.: Automatic Bass Line Transcription from Polyphonic Music, in *Proc. of ICMC 2001*, 2001, pp. 431–434.

[71] Han, Y. and Raphael, C.: Desoloing Monaural Audio Using Mixture Models, in *Proc. of ISMIR 2007*, 2007.

[72] Hori, T., Hori, C. and Minami, Y.: Speech Summarization Using Weighted Finite-State Transducers, in *Proc. of Eurospeech 2003*, 2003, pp. 2817–2820.

[73] HTK: http://htk.eng.cam.ac.uk/.

[74] Hu, N., Dannenberg, R. B. and Tzanetakis, G.: Polyphonic Audio Matching and Alignment for Music Retrieval, in *Proc. of WASPAA 2003*, 2003, pp. 185–188.

[75] Iskandar, D., Wang, Y., Kan, M.-Y. and Li, H.: Syllabic Level Automatic Synchronization of Music Signals and Text Lyrics, in *Proc. of ACM Multimedia 2006*, 2006, pp. 659–662.

[76] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus, in *Proc. of ICSLP'98*, Vol. 7, 1998, pp. 5121–5124.

[77] İzmirli, Ö.: Audio Key Finding Using Low-Dimensional Spaces, in *Proc. of ISMIR 2006*, 2006.

[78] İzmirli, Ö.: Localized Key Finding from Audio Using Non-Negative Matrix Factorization for Segmentation, in *Proc. of ISMIR 2007*, 2007.

[79] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C.: The ICSI Meeting Corpus, in *Proc. of ICASSP 2003*, Vol. 1, 2003, pp. 364–367.

[80] jMIR: http://jmir.sourceforge.net/.

[81] Julius: http://julius.sourceforge.jp/.

[82] Juman: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html.

[83] Kameoka, H.: *Statistical Approach to Multipitch Analysis*, PhD Thesis, The University of Tokyo, 2007.

[84] Kameoka, H., Nishimoto, T. and Sagayama, S.: A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 3(2007), pp. 982–994.

[85] Kashino, K. and Murase, H.: A Sound Source Identification System for Ensemble Music Based on Template Adaptation and Music Stream Extraction, *Speech Communication*, Vol. 27, No. 3–4(1999), pp. 337–349.

[86] Kawahara, T., Lee, C.-H. and Juang, B.-H.: Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 6, No. 6(1998), pp. 558–568.

[87] Kitahara, T.: *Computational Musical Instrument Recognition and Its Application to Content-based Music Information Retrieval*, PhD Thesis, Kyoto University, 2007.

[88] Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Instrogram: Probabilistic Representation of Instrument Existence for Polyphonic Music, *IPSJ Journal* (情報処理学会論文誌), Vol. 48, No. 1(2007), pp. 214–226.

[89] Klapuri, A. and Davy, M.(eds.): *Signal Processing Methods for Music Transcription*, Springer, 2006.

[90] Klapuri, A. P., Eronen, A. J., and Astola, J. T.: Analysis of the Meter of Acoustic Musical Signals, *IEEE Transactions on Speech and Audio Processing*, Vol. 14, No. 1 (2006), pp. 342–355.

[91] Lee, K. F.: *Automatic Speech Recognition: The Development of the Sphinx Recognition System*, Kluwer Academic Publishers, 1989.

[92] Lee, K. and Slaney, M.: Automatic Chord Recognition from Audio Using an HMM with Supervised Learning, in *Proc. of ISMIR 2006*, 2006.

[93] Lee, K. and Slaney, M.: A Unified System for Chord Transcription and Key Extraction Using Hidden Markov Models, in *Proc. of ISMIR 2007*, 2007.

[94] Legetter, C. J. and Woodland, P. C.: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density Hidden Markov Models, *Computer Speech & Language*, Vol. 9, No. 2(1995), pp. 171–186.

[95] Leveau, P., Sodoyer, D. and Daudet, L.: Automatic Instrument Recognition in a Polyphonic Mixture Using Sparse Representations, in *Proc. of ISMIR 2007*, 2007.

[96] Lidy, T., Rauber, A., Pertusa, A. and Iñesta, J. M.: Improving Genre Classification by Combination of Audio and Symbolic Descriptors Using a Transcription System, in *Proc. of ISMIR 2007*, 2007.

[97] Logan, B.: Content-Based Playlist Generation: Exploratory Experiments, in *Proc. of ISMIR 2002*, 2002, pp. 295–296.

[98] Logan, B. and Chu, S.: Music Summarization Using Key Phrases, in *Proc. of ICASSP 2000*, 2000, pp. II–749–752.

[99] Lu, L., Zhang, H.-J. and Li, S.: Content-based Audio Classification and Segmentation by Using Support Vector Machines, *ACM Multimedia Systems Journal*, Vol. 8, No. 6(2003), pp. 482–492.

[100] M2K: http://www.music-ir.org/evaluation/m2k/.

[101] Marolt, M.: Gaussian Mixture Models for Extraction of Melodic Lines from Audio Recordings, in *Proc. of ISMIR 2004*, 2004.

[102] MARSYAS: http://marsyas.sness.net/.

[103] Master, A. S.: *Stereo Music Source Separation Via Bayesian Modeling*, PhD Thesis, Stanford University, 2006.

[104] MATLAB: http://www.mathworks.com/products/matlab/.

[105] McKinney, M. F. and Breebaart, J.: Features for Audio and Music Classification, in *Proc. of ISMIR 2003*, 2003, pp. 151–158.

[106] Mecab: http://mecab.sourceforge.net/.

[107] Mesaros, A., Virtanen, T. and Klapuri, A.: Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods, in *Proc. of ISMIR 2007*, 2007.

[108] Metze, F., Waibel, A., Bett, M., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H. and Zech-

ner, K.: Advances in Automatic Meeting Record Creation and Access, in *Proc. of ICASSP 2001*, Vol. 1, 2001, pp. 601–604.

[109] MIRtoolbox: http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox.

[110] Moreau, A. and Flexer, A.: Drum Transcription in Polyphonic Music Using Non-Negative Matrix Factorisation, in *Proc. of ISMIR 2007*, 2007.

[111] Müller, M. and Clausen, M.: Transposition-Invariant Self-Similarity Matrices, in *Proc. of ISMIR 2007*, 2007.

[112] Nishimura, T., Hashiguchi, H., Takita, J., Zhang, J. X., Goto, M. and Oka, R.: Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming, in *Proc. of ISMIR 2001*, 2001, pp. 211–218.

[113] Ogata, J., Goto, M. and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, in *Proc. of Interspeech 2007*, 2007.

[114] Oliver, N. and Kreger-Stickles, L.: PAPA: Physiology and Purpose-Aware Automatic Playlist Generation, in *Proc. of ISMIR 2006*, 2006.

[115] Opolko, F. and Wapnick, J.: McGill University Master Samples (CDs), 1987.

[116] Paiva, R. P., Mendes, T. and Cardoso, A.: An Auditory Model Based Approach for Melody Detection in Polyphonic Musical Recordings, in *Proc. of International Symposium on Computer Music Modeling and Retrieval*, 2004.

[117] Palmkit: http://palmkit.sourceforge.net/.

[118] Pampalk, E. and Gasser, M.: An Implementation of a Simple Playlist Generator Based on Audio Similarity Measures and User Feedback, in *Proc. of ISMIR 2006*, 2006.

[119] Pampalk, E. and Goto, M.: MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling, in *Proc. of ISMIR 2006*, 2006.

[120] Paulus, J. and Klapuri, A.: Measuring the Similarity of Rhythmic Patterns, in *Proc. of ISMIR 2002*, 2002, pp. 150–156.

[121] Paulus, J. and Klapuri, A.: Combining Temporal and Spectral Features in HMM-Based Drum Transcription, in *Proc. of ISMIR 2007*, 2007.

[122] Pauws, S. and Eggen, B.: PATS: Realization and User Evaluation of an Automatic Playlist Generator, in *Proc. of ISMIR 2002*, 2002, pp. 222–230.

[123] Pauws, S., Verhaegh, W. and Vossen, M.: Fast Generation of Optimal Music Playlists Using Local Search, in *Proc. of ISMIR 2006*, 2006.

[124] Peeters, G.: Chroma-based Estimation of Musical Key from Audio-signal Analysis, in *Proc. of ISMIR 2006*, 2006.

[125] Peeters, G.: Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum-Likelihood Approach, in *Proc. of ISMIR 2007*, 2007.

[126] Peeters, G., Burthe, A. L. and Rodet, X.: Toward Automatic Music Audio Summary Generation from Signal Analysis, in *Proc. of ISMIR 2002*, 2002, pp. 94–100.

[127] Peng, W., Li, T. and Ogihara, M.: Music Clustering with Constraints, in *Proc. of ISMIR 2007*, 2007.

[128] Pikrakis, A., Giannakopoulos, T. and Theodoridis, S.: A Computationally Efficient Speech/Music Discriminator for Radio Recordings, in *Proc. of ISMIR 2006*, 2006.

[129] Pinquier, J., Rouas, J.-L. and Andre-Obrecht, R.: A Fusion Study in Speech/Music Classification, in *Proc. of ICASSP 2003*, Vol. 2, 2003, pp. 17–20.

[130] Pohle, T., Knees, P., Schedl, M. and Widmer, G.: Independent Component Analysis for Music Similarity Computation, in *Proc. of ISMIR 2006*, 2006.

[131] Poliner, G. E., Ellis, D. P. W., Ehmann, A. F., Gomez, E., Streich, S. and Ong, B.: Melody Transcription from Music Audio: Approaches and Evaluation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 4(2007), pp. 1247–1256.

[132] Rabiner, L. and Juang, B.-H.: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[133] Reed, J. and Lee, C.-H.: A Study on Music Genre Classification Based on Universal Acoustic Models, in *Proc. of ISMIR 2006*, 2006.

[134] Reynolds, D. A.: An Overview of Automatic Speaker Recognition Technology, in *Proc. of ICASSP 2002*, 2002, pp. V–4072–4075.

[135] Rhodes, C. and Casey, M.: Algorithms for Determining and Labelling Approximate Hierarchical Self-Similarity, in *Proc. of ISMIR 2007*, 2007.

[136] Roads, C.: *The Computer Music Tutorial*, The MIT Press, 1996.

[137] Rosenfeld, R.: A Maximum Entropy Approach to Adaptive Statistical Language Modeling, *Computer Speech & Language*, Vol. 10, No. 3(1996), pp. 187–228.

[138] RWTH ASR: http://www-i6.informatik.rwth-aachen.de/rwth-asr/.

[139] Ryynänen, M. and Klapuri, A.: Transcription of the Singing Melody in Polyphonic Music, in *Proc. of ISMIR 2006*, 2006, pp. 222–227.

[140] Ryynänen, M. and Klapuri, A.: Automatic Bass Line Transcription from Streaming Polyphonic Audio, in *Proc. of ICASSP 2007*, 2007, pp. IV–1437–1440.

[141] Scheirer, E. and Slaney, M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, in *Proc. of ICASSP'97*, Vol. 2, 1997, pp. 1331–1334.

[142] Scheirer, E. D.: Tempo and Beat Analysis of Acoustic Musical Signals, *J. Acoust. Soc. Am.*, Vol. 103, No. 1(1998), pp. 588–601.

[143] Schmidt, M. N. and Morup, M.: Sparse Non-negative Matrix Factor 2-D Deconvolution for Automatic Transcription of Polyphonic Music, in *Proc. of ICA 2006*, 2006.

[144] Schuller, B., Rigoll, G. and Lang, M.: Hidden Markov Model-based Speech Emotion Recognition, in *Proc. of ICASSP 2003*, Vol. 2, 2003, pp. 1–4.

[145] Schwartz, S., Chow, Y.-L., Kimball, O., Roucos, S., Krasner, M. and Makhoul, J.: Context-dependent Modeling for Acoustic-phonetic of Continuous Speech, in *Proc. of ICASSP'85*, 1985, pp. 1205–1208.

[146] Schwenninger, J., Brueckner, R., Willett, D. and Hennecke, M.: Language Identification in Vocal Music, in *Proc. of ISMIR 2006*, 2006.

[147] Seppänen, J., Eronen, A. and Hiipakka, J.: Joint Beat & Tatum Tracking from Music Signals, in *Proc. of ISMIR 2006*, 2006.

[148] Sheh, A. and Ellis, D. P. W.: Chord Segmentation and Recognition Using EM-Trained Hidden Markov Models, in *Proc. of ISMIR 2003*, 2003.

[149] Shinoda, K. and Watanabe, T.: MDL-based Context-Dependent Subword Modeling for Speech Recognition, *J. Acoust. Soc. Jpn.* (E), Vol. 21, No. 2(2000), pp. 79–86.

[150] Sloboda, T.: Dictionary Learning: Performance through Consistency, in *Proc. of ICASSP'95*, Vol. 1, 1995, pp. 453–456.

[151] Sloboda, T. and Waibel, A.: Dictionary Learning for Spontaneous Speech Recognition, in *Proc. of ICSLP'96*, Vol. 4, 1996, pp. 5121–5124.

[152] Smaragdis, P. and Brown, J. C.: Non-Negative Matrix Factorization for Polyphonic Music Transcription, in *Proc. of WASPAA 2003*, 2003, pp. 177–180.

[153] Song, J., Bae, S. Y. and Yoon, K.: Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System, in *Proc. of ISMIR 2002*, 2002, pp. 133–139.

[154] Sonic Visualiser: http://www.sonicvisualiser.org/.

[155] Sphinx: http://cmusphinx.sourceforge.net/.

[156] SRILM: http://www.speech.sri.com/projects/srilm/.

[157] Steinbiss, V. and Ney, H.: Continuous Speech Dictation - From Theory to Practice, *Computer Speech & Language*, Vol. 13, No. 2(1995), pp. 19–38.

[158] Suzuki, M., Hosoya, T., Ito, A. and Makino, S.: Music Information Retrieval from a Singing Voice Based on Verification of Recognized Hypotheses, in *Proc. of ISMIR 2006*, 2006.

[159] Torres, D., Turnbull, D., Barrington, L. and Lanckriet, G.: Identifying Words That Are Musically Meaningful, in *Proc. of ISMIR 2007*, 2007.

[160] Tranter, S. E. and Reynolds, D. A.: An Overview of Automatic Speaker Diarization Systems, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5(2006), pp. 1557–1565.

[161] Tsai, W.-H. and Wang, H.-M.: Automatic Singer Recognition of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signals, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1(2006), pp. 330–341.

[162] Tsai, W.-H. and Wang, H.-M.: Automatic Identification of the Sung Language in Popular Music Recordings, *Journal of New Music Research*, Vol. 36, No. 2(2007), pp. 105–114.

[163] Tsai, W.-H., Wang, H.-M., Rodgers, D., Cheng, S.-S. and Yu, H.-M.: Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics, in *Proc. of ISMIR 2003*, 2003, pp. 167–173.

[164] Turnbull, D., Barrington, L. and Lanckriet, G.: Modeling Music and Words Using a Multi-Class naïve Bayes Approach, in *Proc. of ISMIR 2006*, 2006.

[165] Turnbull, D., Lanckriet, G., Pampalk, E. and Goto, M.: A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting, in *Proc. of ISMIR 2007*, 2007.

[166] Tzanetakis, G. and Cook, P.: Musical Genre Classification of Audio Signals, *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5(2002), pp. 293–302.

[167] Tzanetakis, G., Jones, R. and McNally, K.: Stereo Panning Features for Classifying Recording Production Style, in *Proc. of ISMIR 2007*, 2007.

[168] Vamp audio analysis plugin system: http://www.vamp-plugins.org/.

[169] van de Par, S., McKinney, M. and Redert, A.: Musical Key Extraction from Audio Using Profile Training, in *Proc. of ISMIR 2006*, 2006.

[170] Varewyck, M. and Martens, J.-P.: Assessment of State-of-The-Art Meter Analysis Systems with an Extended Meter Description Model, in *Proc. of ISMIR 2007*, 2007.

[171] Ververidis, D. and Kotropoulos, C.: Emotional Speech Recognition: Resources, Features, and Methods, *Speech Communication*, Vol. 48, No. 9(2006), pp. 1162–1181.

[172] Watanabe, S., Sako, A. and Nakamura, A.: Automatic Determination of Acoustic Model Topology Using Variational Bayesian Estimation and Clustering for Large Vocabulary Continuous Speech Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 14, No. 3(2006), pp. 855–872.

[173] WaveSurfer: http://www.speech.kth.se/wavesurfer/.

[174] Weka: http://www.cs.waikato.ac.nz/ml/weka/.

[175] Whitman, B. and Smaragdis, P.: Combining Musical and Cultural Features for Intelligent Style Detection, in *Proc. of ISMIR 2002*, 2002, pp. 47–52.

[176] Womack, B. D. and Hansen, J. H. L.: N-Channel Hidden Markov Models for Combined

Stressed Speech Classification and Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 7, No. 6(1999), pp. 668–677.
[177] Woodland, P. C., Gales, M. J. F., Pye, D. and Young, S. J.: Broadcast News Transcription Using HTK, in *Proc. of ICASSP'97*, Vol. 2, 1997, pp. 719–722.
[178] Woodruff, J., Pardo, B. and Dannenberg, R.: Remixing Stereo Music with Score-Informed Source Separation, in *Proc. of ISMIR 2006*, 2006.
[179] Yoshii, K., Goto, M. and Okuno, H. G.: Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates with Harmonic Structure Suppression, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 1(2007), pp. 333–345.
[180] Yoshioka, T., Kitahara, T., Komatani, K., Ogata, T. and Okuno, H. G.: Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries, in *Proc. of ISMIR 2004*, 2004.
[181] Yu, H., Clark, C., Malkin, R. and Waibel, A.: Experiments in Automatic Meeting Transcription Using JRTk, in *Proc. of ICASSP'98*, Vol. 2, 1998, pp. 921–924.
[182] Zils, A., Pachet, F., Delerue, O. and Gouyon, F.: Automatic Extraction of Drum Tracks from Polyphonic Music Signals, in *Proc. of WEDELMUSIC 2002*, 2002, pp. 179–183.
[183] 中山一郎: 日本語を歌・唄・謡う, 日本音響学会誌, Vol. 59, No. 11(2003), pp. 688–693.
[184] 荒木雅弘: フリーソフトでつくる音声認識システム, 森北出版, 2007.
[185] 今井亨, リチャードシュワルツ, 小林彰夫, 安藤彰男: 話題混合モデルによる放送ニュースからの話題抽出, 信学論 (D-II), Vol. J81-D-II, No. 9(1998), pp. 1955–1964.
[186] 中村匡伸, 岩野公司, 古井貞熙: 日本語話し言葉コーパスを用いた話し言葉音声の音響的特徴の分析, 情処研報音声言語情報処理 2004-SLP-53-2, 2004.
[187] 広瀬啓吉, 阪田真弓: 対話音声と朗読音声の韻律的特徴の比較, 信学論 (D-II), Vol. 79, No. 12(1996), pp. 2154–2162.
[188] 岩野公司, 広瀬啓吉: 語彙制約なし音声認識へのアクセント句境界検出の統合, 信学論 (D-II), Vol. 83, No. 10(2000), pp. 1977–1985.
[189] 高木幸一, 桜井直之, 岩崎淳, 古井貞熙: ニュース音声を対象とした言語モデルと話題抽出の検討, 信学技報 音声 SP98-33, 1998.
[190] 大石康智, 後藤真孝, 伊藤克亘, 武田一哉: スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別, 情報処理学会論文誌, Vol. 47, No. 6(2006), pp. 1822–1830.
[191] 尾関弘尚, 鎌田貴幸, 後藤真孝, 速水悟: 歌声の歌詞認識における音高の影響について, 音講論集 (秋)1-1-1, 2003.
[192] 糸山克寿, 後藤真孝, 駒谷和範, 尾形哲也, 奥乃博: 楽譜情報を援用した多重奏音楽音響信号の音源分離と調波・非調波統合モデルの制約付パラメータ推定の同時実現, 情報処理学会論文誌, Vol. 49, No. 3(2008), pp. 1465–1479.
[193] 深田俊明, 匂坂芳典: 発音ネットワークに基づく発音辞書の自動生成, 信学論 (D-II), Vol. J80-D-II, No. 10(1997), pp. 2626–2635.
[194] 佐古淳, 滝口哲也, 有木康雄: AdaBoost を用いたシステムへの問い合わせと雑談の判別, 信学技報 音声 2006-SP-88, (2006), pp. 19–24.
[195] 緒方淳, 有木康雄: 日本語話し言葉音声認識のための音節に基づく音響モデリング, 信学論 (D-II), Vol. J86-D-II, No. 11(2003), pp. 1523–1530.
[196] 緒方淳, 後藤真孝, 江渡浩一郎: PodCastle: ポッドキャストをテキストで検索, 閲覧, 編集できるソーシャルアノテーションシステム, WISS 2006 論文集, 2006, pp. 53–58.
[197] 緒方淳, 後藤真孝, 江渡浩一郎: PodCastle の実現: Web 2.0 に基づく音声認識性能の向上について, 情処研報 音声言語情報処理 2007-SLP-65-8, 2007, pp. 41–46.
[198] 小林彰夫, 今井亨, 安藤彰男, 中林克己: ニュース音声認識のための時期依存言語モデル, 情報処理学会論文誌, Vol. 40, No. 4(1999), pp. 1421–1429.
[199] 河口信夫, 松原茂樹, 山口由紀子, 武田一哉, 板倉文忠: CIAIR 実走行車内音声データベース, 情処研報 音声言語情報処理 2003-SLP-49-24, 2003.
[200] 後藤真孝: 音楽音響信号を対象としたリアルタイムビートトラッキングに関する研究, 博士論文, 早稲田大学, 1998.
[201] 後藤真孝, 齋藤毅, 中野倫靖, 藤原弘将: 歌声情報処理の最近の研究, 日本音響学会誌, Vol. 64, No. 10(2008), pp. 616–623.
[202] 後藤真孝, 平田圭二: 音楽情報処理の最近の研究, 日本音響学会誌, Vol. 60, No. 11(2004), pp. 675–681.
[203] 後藤真孝, 北山広治, 伊藤克亘, 小林哲則: 音声スポッタ: 人間同士の会話中に音声認識が利用可能な音声入力インタフェース, 情報処理学会論文誌, Vol. 48, No. 3(2007), pp. 1274–1283.
[204] 後藤真孝, 緒方淳, 江渡浩一郎: PodCastle の提案: 音声認識研究 2.0 を目指して, 情処研報 音声言語情報処理 2007-SLP-65-7, 2007, pp. 35–40.
[205] 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(2004), pp. 728–738.
[206] 藤江真也, 江尻康, 菊池英明, 小林哲則: 肯定的/否定的発話態度の認識とその音声対話システムへの応用, 信学論 (D-II), Vol. J88-D-II, No. 3(2005), pp. 489–498.
[207] 中川聖一: 音声認識研究の動向, 信学論 (D-II), Vol. J83-D-II, No. 2(2000), pp. 433–457.
[208] 中川聖一, 花井建豪, 山本一公, 峯松信明: HMM に基づく音声認識のための音節モデルと triphone モデルの比較, 信学論 (D-II), Vol. J83-D-II, No. 6(2000), pp. 1412–1421.
[209] 工藤拓, 山本薫, 松本裕治: Conditional Random Fields を用いた日本語形態素解析, 情処研報 自然言語処理 2004-NL-161-13, 2004, pp. 89–96.
[210] 河原達也, 武田一哉, 伊藤克亘, 李晃伸, 鹿野清宏, 山田篤: 連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要, 情処研報 音声言語情報処理

2003-SLP-49-57, 2003, pp. 325–330.
[211] 松井知子: 音声による個人認証: 話者認識技術の研究動向, 日本音響学会誌, Vol. 63, No. 12(2007), pp. 738–743.
[212] 松井知子, 内藤正樹, 中村篤, 匂坂芳典: 地域や年齢的な広がりを考慮した大規模な日本語音声データベース, 音講論集 (秋)3-Q-26, 1999.
[213] 堀智織, 古井貞熙: 単語抽出による音声要約文生成法とその評価, 信学論 (D-II), Vol. J85-DII, No. 2(2002), pp. 200–209.
[214] 古井貞熙, 前川喜久雄, 伊佐原均: 科学技術振興調整費開放的融合研究推進制度－大規模コーパスに基づく「話し言葉工学」の構築－, 日本音響学会誌, Vol. 56, No. 11(2000), pp. 752–755.
[215] 藤崎博也: 音声の韻律的特徴における言語的・パラ言語的・非言語的情報の表出, 信学技報 ヒューマンコミュニケーション 1994-HC-94-217, 1994, pp. 1–8.
[216] 松原勇介, 緒方淳, 後藤真孝: ポッドキャスト音声認識の性能向上手法:集合知によって更新される Web キーワードを活用した言語モデリング, 情処研報 自然言語処理 2008-NL-185-6, (2008), pp. 39–44.
[217] 北出祐, 南條浩輝, 河原達也, 奥乃博: 談話標識と話題語に基づく統計的尺度による講演からの重要文抽出, 情処研報音声言語情報処理 2003-SLP-46-2, 2003.
[218] 秋田裕哉, 河原達也: 話し言葉音声認識のための汎用的な統計的発音変動モデル, 信学論 (D-II), Vol. J86-D-II, No. 9(2005), pp. 1780–1789.
[219] 松本裕治: 形態素解析システム「茶筌」, 情報処理 (情報処理学会誌), Vol. 41, No. 11(2000), pp. 1208–1214.
[220] 中野倫靖, 後藤真孝, 平賀譲: 楽譜情報を用いない歌唱力自動評価手法, 情報処理学会論文誌, Vol. 48, No. 1(2007), pp. 227–236.