修 士 論 文

雑音環境下における音源分離を認識規範とした音声認識に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

羽二生 篤

2004 年 3 月

修 士 論 文

雑音環境下における音源分離を認識規範とした音声認識に関する研究

指導教官　赤木正人 教授

審査委員主査　赤木正人 教授
審査委員　党建武 助教授
審査委員　下平博 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

210070 羽二生 篤

提出年月: 2004 年 2 月

概要

本論文では、分離過程で目的音として妥当な音を分離しているかどうか、最終的な分離音が目的音として妥当であるかどうかを判断することにより、入力音中の目的音を認識する音源分離を認識の規範とした音声認識の手法を提案する。

妥当な振る舞いをする音源分離手法において、分離過程で目的音と思われる音を知識を用いて積極的に分離する過程を経て、最終的な分離音が目的音らしといえるのであれば、入力音中は目的音が存在したといえる。このような、音源分離システムを用いて、分離の過程と結果の妥当性を判断することができれば、入力音中の目的音を認識できると考えられる。このコンセプトにもとづき、音源分離処理と認識処理が融合した、音源分離を認識規範とする音声認識モデルを構築した。

提案した手法の有効性を示すために、白色雑音と単母音が混合した音から目的音声を認識するシミュレーションを行った。これにより、定常な白色雑音と単母音が空間的に加算された状況において本手法が SNR が -10 dB まで有効である事を確認した。

目次

図目次

表目次

第 1 章　序論

1.1　はじめに

現在、機械に対し利用者の意思を伝えるためにはキーボードやマウス、タッチパネルのようなデバイスが広く使われている。これらの入力デバイスは事前に利用者がその使用方法を学習する必要がある。特にキーボードなどはそのキー数の多さなどから高齢者などがコンピュータの利用を敬遠する傾向にある。さらに、入力には手全体や指先を利用するために障害者などにとっては高いハードルとなることがある。このようなことから、現在の機械への入力手段はデジタルデバイドが生じる一要因とされている [郵政 99]。機械への入力手段として、我々が日常用いている音声を利用することができれば、入力に際して特別な学習がほとんど不要になり、より多くの人々がコンピュータを利用することが可能になる。また、多くの障害者に対してもコンピュータの利用に門戸を開くこととなる。さらに、健常者にとっても四肢や視線を拘束されずコンピュータへの入力が行えることから、コンピュータの利用時に行動が拘束されなくなるという利点もある。コンピュータでのユーザーインターフェースでの利用のほかに、自動翻訳電話、字幕放送の自動化、ひいてはロボットとの言葉による自然なやり取りなどの技術応用が考えられる。このような数多くの利点があるため、機械で音声を認識する研究はコンピュータが開発されて間もない 1950 年代から行われている [古井 85, Dav52]。現在までに行われている音声認識手法としては、DP (Dynamic Programing) マッチング法、HMM (Hidden Markov Model)、ニューラルネットワークを用いたものなどがある。これらは、雑音がなく一人の話者の音声のみが存在する理想的な環境においてある程度高い認識率を誇り、次の段階として雑音環境下や実環境での認識率の向上が課題となっている。

我々が生活する実環境には、周囲に様々な音源が存在する。各音源から発生する音の周波数成分は、音源ごとに異なっている。さらに、各音源はそのほとんどが独立に存在しているものであって、個々の音源から音が発生するタイミングもまた独立である。このため実環境に存在する音は、時間的に重なりを持つと同時に周波数的にも重なりを持ち、しかもどのよ

うな重なりとなるかをあらかじめ予測することは現実的に難しい。したがって、実環境において音声を認識するということは、目的音である音声とそれ以外の周囲から発生した背景音とが周波数成分も時間成分もともに予測不可能な重なりを持つ混合音から目的音の内容を認識するということである。このような雑音環境下や実環境下での認識手法としては理想的な環境で用いられている手法をもとに、スペクトルサブトラクションやマイクロホンアレイのような雑音抑圧や音声強調を音声認識の前処理として用いるもの [Flo94, 武田 97, 金田 97] や HMM 合成法のように音響モデルに雑音 HMM を重畳させ音響モデルを変形するもの [Min92, Tak01, Vas97] が主に研究されている。しかし、現段階では未知の雑音や非定常、特に突発的な雑音が存在する状況において精度よく認識を行うことができず、未だに実環境での使用に耐えうる音声認識システムはない。

一方、我々人間は、実環境中において、いとも容易く人の声を聞き分け、その内容を理解することができる。例えば、人間は雑踏の中においても隣の人と会話を続けることができ、走行中の車中においても同乗者と会話を行うことができる。図 1.1 に示したのは、雑音環境下での人間と機械の音声認識における誤聴率を示したものである [Lip97]。機械の音声認識は雑音の影響を受けて誤聴率が上昇するのに対して、人間は雑音の影響をほとんど受けていないことがわかる。このように雑音が存在する状況下でも特定の音に注意を向けて、その音を聞き分けることができる能力は “カクテルパーティ効果” として知られている。このように人間の聴覚は非常に優れた能力を持っているため、人間の聴覚モダリティを解明することは機械の音声認識において何らかの手がかりになるのではないかと考えられている [中川 00]。その一方で “McGurk 効果” や “腹話術師効果”、またその逆に近年発見された聴覚モダリティが視覚モダリティに影響を与える効果 [Shi01] などから伺えるように単一の知覚モダリティが現実世界で生じている物理的事象を常に正確に表現しているとは限らず、聴覚モダリティもその例外ではない。このことは、図 1.1 においてクリーンな環境においても人間の誤聴率が 0 % でないことからもわかる。しかし、人間の誤聴率がクリーンな環境下で 0 % でないからといって、機械による音声認識の手本として人間の聴覚がなり得ないわけではない。人間の聴覚はおよそ 5 億年前のカンブリア紀に起きたとされる種の爆発から自然淘汰という非常に強力なフィルタをくぐり抜けてきて現在に至る [小田 02]。それゆえ、人間の聴覚は決して闇雲な解を導きだしている訳ではなく、その生存に十分必要な確度を持っている妥当な結果を導きだしていると考えられる。本研究では雑音環境中でも妥当な解を導きだす人間の聴覚モダリティに注目し、これを発想の原点として雑音環境下における音声認識を試みる。

なお、「音声認識」という言葉には音韻情報などの言語情報の機械による自動認識という狭義の意味と言語情報に加え、発話者の個人性をも認識する話者認識を含めた広義の意味があ

る。本研究で用いる「音声認識」とは狭義の意味で用いており、以降も同様の意味で用いる。

1.2 研究の背景

1.2.1 前処理による雑音環境への対応

既存の音声認識手法、特に HMM を用いた認識手法は雑音が存在しない環境下では、図 1.2 に示すように、ある程度低い誤聴率を誇る。このような既存の認識手法に雑音除去のような前処理を加えることにより雑音環境に適応する手法が研究されている。雑音除去の手法としてはマイクロホンを 1 つ用いるものとそれ以上用いるものとに大きく分けられる。マイクロホンを 1 つ用いるものは、ヘッドセットマイクロホンによるものは別として、周波数領域における信号処理が中心であり、マイクロホンを 2 つ以上用いるものは時間領域での信号処理が中心となる。

マイクロホンを 1 つ用いる手法としてはスペクトルサブトラクション法があげられる。スペクトルサブトラクション法は時刻 t における観測信号 $y(t)$ は音声信号 $s(t)$ と雑音信号 $n(t)$ の線形和で表現できると仮定し、

$$y(t) = s(t) + n(t) \tag{1.1}$$

と表す。このとき $s(t)$ と $n(t)$ が独立であれば、フーリエ変換により式 (1.1) は、

$$\begin{aligned} Y(f) &= S(f) + N(f) \\ S(f) &= Y(f) - N(f) \end{aligned} \tag{1.2}$$

となる。ただし、$Y(f) = \mathscr{F}[y(t)]$、$S(f) = \mathscr{F}[s(t)]$、$N(f) = \mathscr{F}[n(t)]$ である。よって、周波数領域において $Y(f)$ から $N(f)$ を引き去ることにより、$S(f)$ を得ることができることになる。正確な $N(f)$ を知ることは難しいが、音声信号が存在しない時に $\overline{N}(f)$ を推定することにより雑音を除去する。

マイクロホンを 2 つ以上用いるものとしてはマイクロホンアレイを用いたものがあげられる。マイクロホンアレイ上の各マイクロホンから得られる信号から特定の方向の音だけを抽出するには、ビームフォーミングや独立成分分析といった手法が用いられる。ビームフォーミングで多く用いられるのは、音源と各マイクロホンからの距離の差により生じる時間差を利用した遅延和アレイがある。ビームフォーミングは $N+1$ 本のマイクロホンにより生じる N 個の音響的死角を雑音に向けることにより雑音を除去する手法である。独立成分分析の場合は、N 本のマイクロホンから観測される混合信号 $\boldsymbol{y}(t) = [y_1(t) y_2(t) \ldots y_N(t)]^{\mathrm{t}}$ は M 個の独立な音源から生じた音 $\boldsymbol{s}(t) = [s_1(t) s_2(t) \ldots s_M(t)]^{\mathrm{t}}$ が線形な空間的混合により生じたものと

仮定し、

$$\boldsymbol{y}(t) = \mathbf{X} \cdot \boldsymbol{s}(t) \tag{1.3}$$

と表す。ただし、$\mathbf{X}$ は $N \times M$ からなる音源からの信号と観測信号の関係を表す混合行列である。この場合、各音源からの信号が独立で $N \geqq M$ となるような条件下を仮定し混合行列の逆行列 $\overline{\mathbf{X}}^{-1}$ を推定することにより、観測信号から各音源で生じた音を推定する。

スペクトルサブトラクションは定常雑音を想定しているために、非定常雑音や突発雑音への対応が困難であるという問題点がある。また、マイクロホンアレイを用いた手法はマイクロホンの数が音源の数と同数かまたはそれ以上である必要があるが、実環境においては音源の数は未知であるという点とこの手法単体ではいずれの分離音が目的音であるかを判断することは難しいという問題点がある。

1.2.2 音響モデルの変形による雑音環境への対応

定常雑音でさえそのスペクトルは決して一定ではない。このような観点から雑音も HMM を用いて表現し、さらにその雑音 HMM を音響モデルに重畳し音響モデルそのものを変形することにより雑音環境下に対応する HMM 合成法が音響モデルを変形する例としてあげられる。近年では特徴パラメータとして対数パワースペクトルやケプストラムが用いられるが、これらの領域では雑音が重畳された音声の特徴パラメータを音声の特徴パラメータと雑音の特徴パラメータの線形和で表現することができない。そこで図 1.3 に示すように、それぞれの特徴パラメータを一旦スペクトル領域に変換して、スペクトル領域でそれぞれのパラメータを重畳し、再び対数パワースペクトルやケプストラムの領域に戻すということで音響モデルの変形を行う [滝口 96]。

HMM 合成法の場合は非定常雑音に対応することが可能である。しかし、雑音も HMM を用いて表現しているために雑音に関しても事前に何らかの学習が必要となる。このため、雑音の性質が突然変化したり、突発的な雑音や想定していない雑音が生じた場合に認識率が低下するという問題点がある。

1.3 研究の目的

前節までに述べたように、既存の手法により実環境での使用に耐えうるような音声システムは現時点では存在しない。そこで本研究では、人間の聴覚に着目し、それに基づき音声認識を行うことを試みる。

人間、機械いずれの音声認識においても、その入力は耳またはマイクにおいて空気の振動

量として与えられる。そして、その時間的変化に基づいて、混合音の中の目的音に対して音声認識処理を行うことになる。音源から生じた音は空気中を伝搬し耳やマイクに到達した時点では、音源位置や音源数、各音源の混合の仕方などの多くの情報が失われた状態となっている。欠落した情報量に対して与えられた情報量はきわめて少ないことから、混合音から目的音を一切の仮定や制約条件なしに一意に決定することは数理工学的に考えて不可能である。

人間は、聴覚情景解析 [Bre90, Bre93] により目的音の妥当な分離結果を導きだしているとされている。聴覚情景解析では 1 つの音源から発生した音の物理的な性質を制約条件として、混合音の中から目的音を分凝する。人間は、このような妥当な音源分離を用いて音声認識を行っていることになる。ボトムアップ的処理である聴覚情景解析とトップダウン的処理である音声認識は人間の知覚の観点から全く独立したものとは考えにくいとされている [Bre94]。したがって、人間の聴覚をもとに音声認識を行おうとした場合には、前処理による雑音環境への適応のように雑音抑圧や音源分離のような処理と認識の処理が完全に独立したようなモデルは不適切であり、音響モデルを変形するような処理も人間の聴覚からかけ離れたものである。そこで本研究では、人間の聴覚が行っている処理をヒントにし、ボトムアップ的な処理である音源分離とトップダウン的な認識が融合したモデルを提案する。さらに、人間が聞きたい音を積極的に選択する “聞き耳” に相当する処理を取り入れ目的音に関する知識を積極的に用いて音源分離行い、このような分離により妥当と思われる分離結果が得られた場合には、入力音中に目的音が存在したと考えるのが妥当であるという立場に立ち、音源分離を認識の規範とした音声認識手法を提案する。そして、この手法により雑音環境下において音声認識を行うことを最終的な目標とし、本論文では今回提案する手法の有効性について検討することを目的とする。

1.4 本論文の構成

本論文は 5 章から構成される。第 1 章は本論文で扱う雑音環境下における音声認識に関する研究の背景と問題点を明らかにし、本論文の特色と目的を示す。

第 2 章では、本論文で扱う状況と問題を設定する。次いで、それに基づく問題の解決法について基本的なコンセプトを示し、本研究で提案する音声認識モデルについて説明する。また、提案モデルの処理の概要についても述べる。

第 3 章では、第 2 章で示した基本原理に基づいたモデルの実装に関して説明する。

第 4 章では、提案手法の評価のために実験の条件と結果、提案手法の有効性を検討した結果を示す。

第 5 章では、本論文の内容を要約し、今後の課題について述べる。

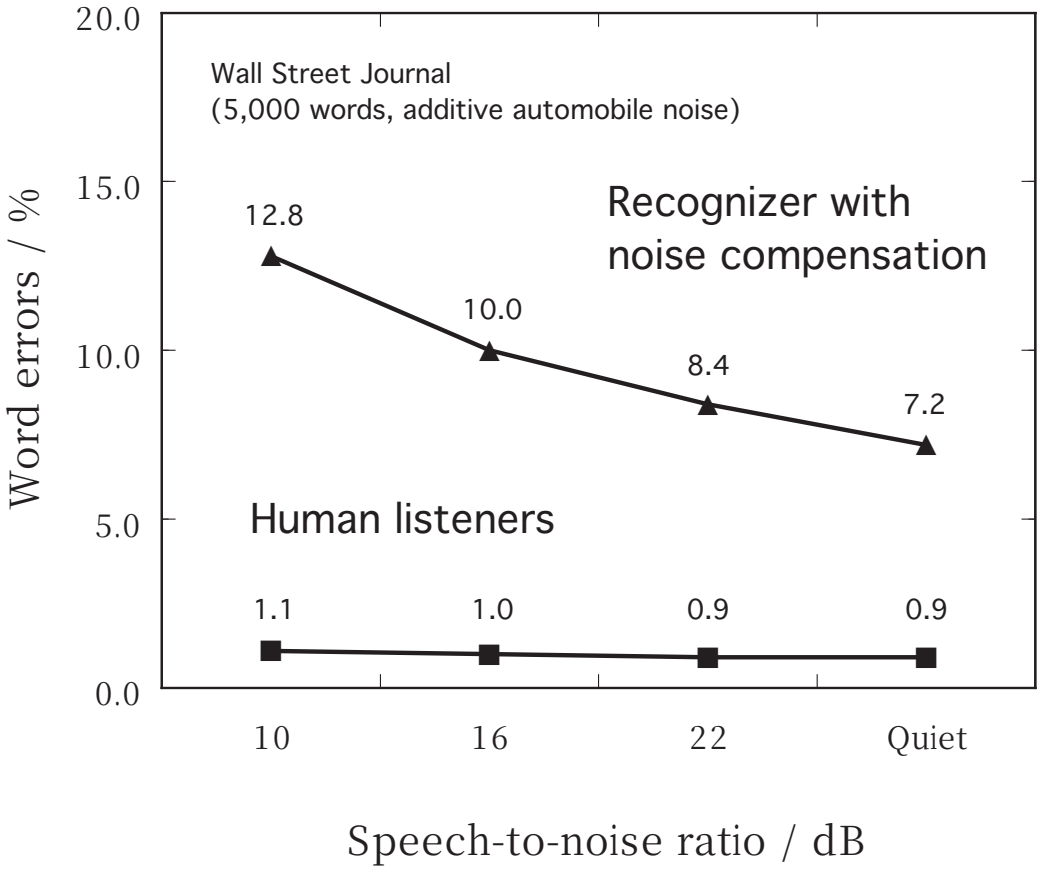


図 1.1 雑音環境下における人間と機械の音声認識における誤聴率 [Lip97]

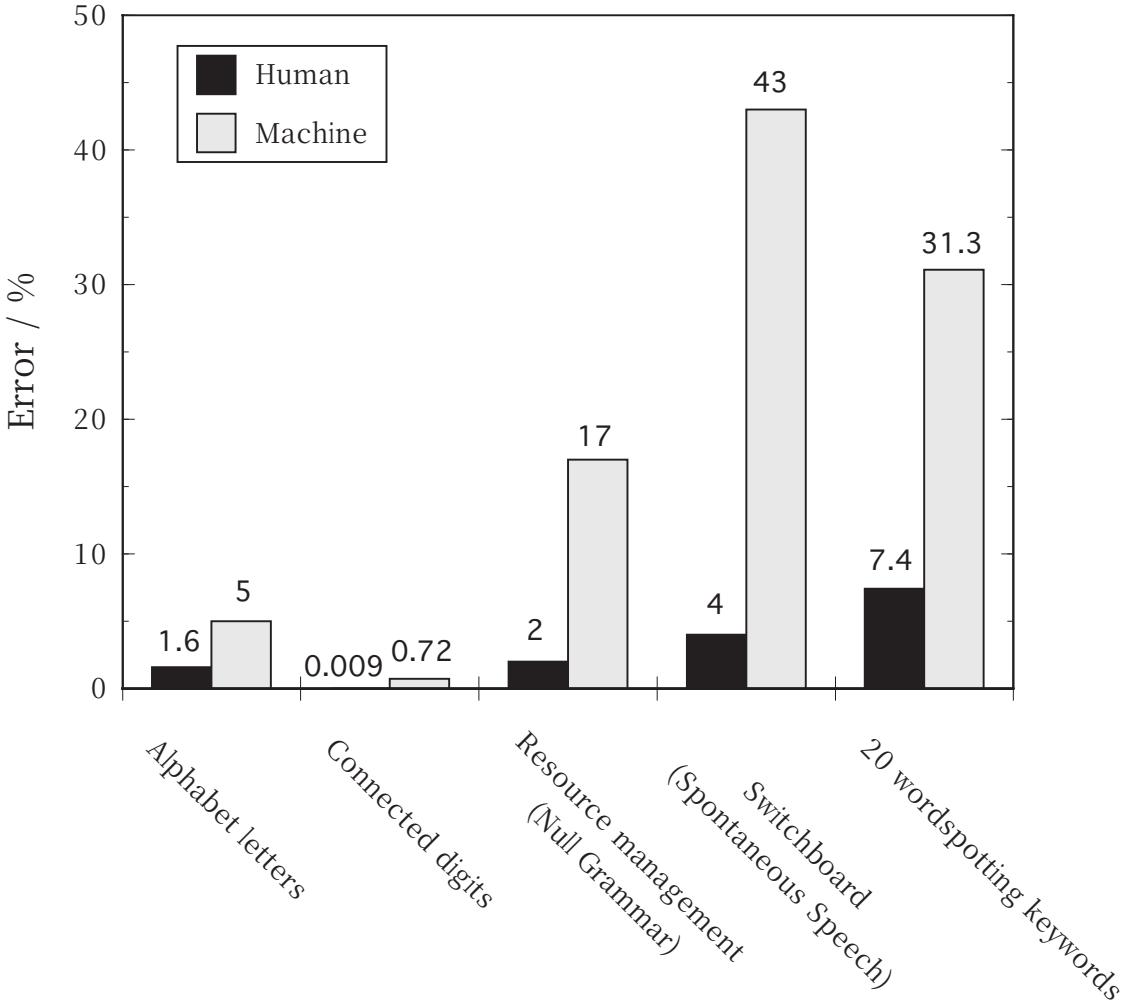


図 1.2 HMM を用いた音声認識システムと人間の誤聴率 [Lip97]

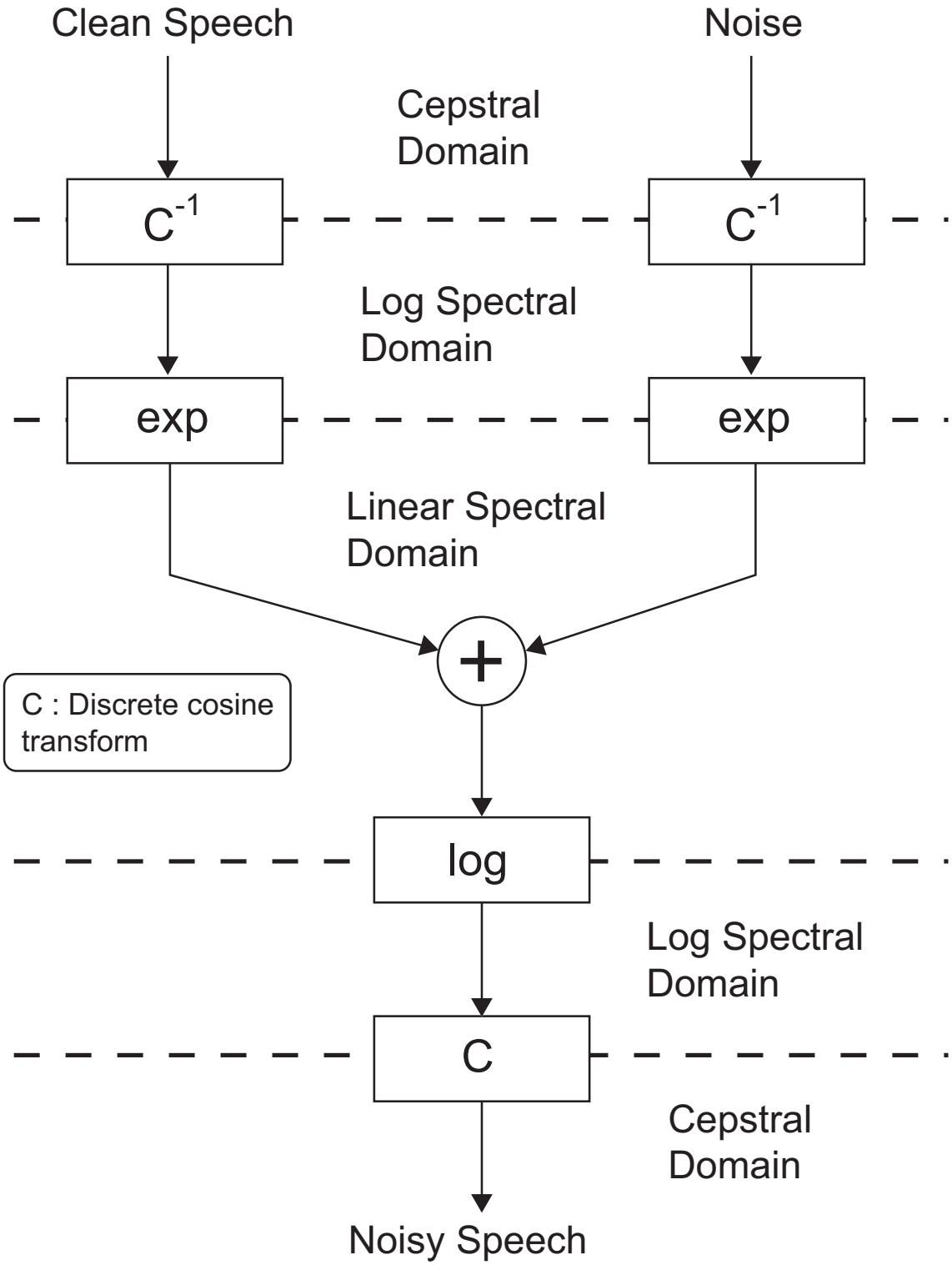


図 1.3 HMM 合成法の処理の流れ

第 2 章　基本原理

本章では、本論文で扱う状況と問題を設定する。次いで、それに基づく問題の解決法について基本的なコンセプトを示し、本研究で提案する音声認識モデルについて説明する。また、提案モデルの処理の概要についても述べる。

2.1　問題設定

前章で述べた理由のほかにも、実環境で音声認識を難しくする要因がある。その要因としては、大きく分けて i) 空間的要因、ii) 電磁気的要因、iii) 話者による要因、があげられる。空間的要因としては、雑音、残響、反射、音源の移動などが考えられる。電磁気的要因としては、マイクロホンから記録機器までの伝送路内での信号の歪み、反射、電磁気的な外来ノイズなどがあげられる。また、録音機器の特性の変動もここに含まれる。話者による要因は、同時発話者の数、感情や Lombard 効果などの内的もしくは外的な原因による発話速度や発声音圧、基本周波数 (F0)、フォルマント周波数の変動などがあげられる [渡辺 96]。これらの要因のうち、システム側で制御不可能で予測困難、かつ、変動が大きい要因は、空間的要因と話者による要因である。

本研究では、目的話者以外に雑音のみが存在する音環境を雑音環境と設定し、その他の要因は考慮しない。この雑音環境下で、本研究が解決しなくてはならない問題は、

1. 目的音と背景音が時間的にも周波数的にも予測不可能な重なりを持つ混合音から目的音を分離する
2. 分離音の内容を認識する

という 2 点となる。

2.2 音源分離を認識規範とした音声認識

本研究では、上記問題を以下のコンセプトに基づき解決することを目指す。

目的音に関する知識を用いて何を分離しているのかが把握できる音源分離システムを考える。このシステムは入力音中に目的音が存在する場合には目的音を分離し、入力音中に目的音が存在しないときは目的音とは異なる音を分離するような妥当な動作をする物だと仮定する。このとき、分離過程で目的音と思われる音を分離するような過程を経て、分離音が目的音として妥当であるといるのであれば、入力音中は目的音が存在したといえる。なぜならば、目的音として妥当な分離経過と結果であれば、その出所は入力音中以外に考えられないからである。例えば、/a/と白色雑音が混合した音を入力音として与え、目的音を/a/として分離を行った場合には/a/が分離される (図 2.1(a))。上記のようなことがこの分離でいえるのであれば、目的音の/a/が入力音中に存在したから目的音が分離されたことになる。同様に、/e/と白色雑音の混合音を入力、目的音を/a/としたときには、/a/とは異なるものが分離される (図 2.1(b))。これは、音声は存在するが、目的音と異なる物であり、そのため目的音とは異なるものが分離された事になる。また、音声が存在せず白色雑音のみを入力して、目的音を/a/としたときには何も分離されない (図 2.1(c))。これもまた、入力音中に目的音が存在しないために何も分離されなかったことになる。

このような、音源分離システムを用いて、分離過程で目的音として妥当な音を分離するような過程を経ているかどうか、最終的な分離音が目的音として妥当であるかどうかを判断することができれば、入力音中に目的音が存在しているのかどうかを判断することができる。本研究では、このような音源分離を認識の規範とするような音声認識の手法を提案し、これにより設定した問題を解決する。

2.3 認識の規範となりうる音源分離手法

音源分離問題では、分離する “音源” として物理的な音源と知覚的な音源の 2 つが考えられる。ここで、物理的な音源とは実際に音を発する物体のことであり、知覚的な音源とは人間が聴感上一つの音と知覚する音のグループを指す [柏野 93]。物理的な音源を分離する手法が、マイクロホンアレイや独立成分分析のような情報理論に基づくブラインド音源分離である。もう一方の、知覚的な音源を分離する手法が、人間の生理学的知見に基づくもの [Bro94] や心理学的知見に基づく手法 [Ell94] のような聴覚情景解析に基づく音源分離である [Coo01]。

ブラインド音源分離は、1.2.1 節で説明したように、式 (1.3) において、混合行列の逆行列を推定し、これにより個々の物理的な音源を導出する物である。この手法を認識の規範として用いることを考えると、分離されたそれぞれの音の中でいずれの音が目的音であるか判断することがこの手法単体では困難である点、情報理論に基づく解法であるため分離の過程を議論することが困難であるという点から、認識の規範として用いるには不向きであると考えられる。一方、聴覚情景解析に基づく音源分離は、Bregman が示した聴覚情景解析の制約条件 [Bre90, Bre93] をもとに知覚的な音源を分離する物である。この手法は制約条件を用いてボトムアップ的に音をグルーピングしていくことが行われている。このボトムアップ処理は音源分離の過程と見なすことが可能であるが、最終的なグルーピングが終了した段階で目的音を判断することになるため、このままでは分離の過程で妥当な分離を行っているのかどうかを判断することは難しい。しかし、この分離過程において、知識を導入して積極的に目的音を分離するような手法を付加することで分離過程が妥当であるかを判断できると考えられる。そこで、本研究では、聴覚情景解析に基づく音源分離手法に対して、知識を用いて積極的に目的音を分離するような手法を加えたものが認識の規範となりうる音源分離手法であるとする。

2.4 本研究のモデル

本研究では、前節までをふまえ、以下のようなモデルにより 2.1 節のような問題を解くことを目指す。本研究の大きな骨組みとなるものは、窪らにより提案された楽器音を対象とした音源分離モデル [窪 02] である。このモデルは、鵜木らが提案した心理学的知見に基づく音源分離モデル [鵜木 99] に対して目的音の知識を利用して分離を行うようにしたものである。このモデルは、前節で述べた聴覚情景解析に基づく音源分離手法に対して知識を導入して積極的に目的音を分離するモデルとなっている。

この窪らのモデルは、複数楽器音の中から目的楽器音を分離することを目指した物で、本研究が想定している状況を対象とはしていない。よって、直接窪らのモデルを本研究で用いることは困難であることから、本研究との前提条件の違いを考慮したモデルを構築することとした。図 2.2 にモデルの概要を示す。本研究のモデルは、1) 信号解析部 (Signal analyzer)、2) 基本周波数 (F0) 推定部 (F0 estimation)、3) 知識制御部 (Knowledge manager)、4) 波形分離部 (Segregation block)、5) 認識部 (Recognition part) という 5 つの処理に大別される。各処理の概要については、以下ようである。

1. 信号解析部 (Signal analyzer) : 入力音を時間と周波数領域の表現に変換する。
2. F0 推定部 (F0 estimation) : 信号解析部の結果を受けて入力音中の F0 を推定する。

3. 知識制御部 (Knowledge manager) : 目的音として指定された音声に関する知識を各処理部の要求に従い提供する。
4. 波形分離部 (Segregation block) : 知識を積極的に用いて目的音として妥当な音声を分離する。
5. 認識部 (Recognition part) : 波形分離部での分離過程と分離音の時間と周波数に関する表現と知識により、最終的な認識を行う。

本研究では、認識の規範として音源分離を用いるのであって、混合音から目的音を分離することが目的ではない。そのため、知覚的な音源を分離する音源分離手法に存在するグルーピングや音声の再合成のような処理は存在しない。

2.5 本モデルによる処理の概要

本モデルは、図 2.2 に示したように、ボトムアップ的な処理と積極的に知識を用いて目的音を選択する音源分離処理と音声認識処理が融合した形をとっている。

2.5.1 音源分離処理

音源分離部での処理は、入力音を目的音と背景音の 2 つの波形を分離するという二波形分離問題に帰着され、セグメントごとに処理を行う。そこで、音源分離処理は、二波形分離問題を処理した鵜木らのモデルを基本に分離を行う。

入力信号は、最初に信号解析部において時間領域の波形が振幅 $S_k(t)$、$(1 \leqq k \leqq K)$ と位相 $f_k(t)$、$(1 \leqq k \leqq K)$ それぞれが時間と周波数領域の表現に変換される。振幅 $S_k(t)$ から F0 推定部において入力音の F0 を推定し調波関係にある周波数を算出する。波形分離部では信号解析部からの $S_k(t)$ と $f_k(t)$、F0 推定部からの調波の周波数、さらに、知識制御部に必要とされる知識を要求して、Bregman の 4 つの発見的規則に基づく立ち上がりたち下がりの同期、調波関係、漸近的変化、振幅包絡間の相関に加え、知識と周波数領域での相関を用いて目的音を分離する。

具体的な処理の流れを図 2.3 に示す。まず始めに、$S_k(t)$ は、F0 推定部で推定された調波の周波数と立ち上がりと立ち下がりに関する規則に基づいて調波関係の周波数成分のみが分離される。次いで、漸近的変化に関する規則により調波成分のみ分離された $S_\ell(t)$、(ℓ は調波位置のチャンネルを示し、$1 < \ell < K$) から推定された分離音の振幅の時間微分 $C_{\ell,R}$ において $C_{\ell,R}$ の推定点とその誤差からいくつかの $C_{\ell,R}$ の候補点を求める。m 番目 $(m \in \ell)$ のチャンネルで $C_{m,R}$ を決定するとき、時間領域と周波数領域で分離音の振幅 $A_\ell(t)$ に関して相関を

とるために、$C_{\ell,R}$ から $A_\ell(t)$ を算出する。このとき、既に m チャンネルより低周波数のチャンネル $(\ell < m)$ では $C_{\ell,R}$ の推定が済んでいるが、m チャンネルより高周波数のチャンネル $(m < \ell \leqq K)$ では推定が済んでいないので、前段階で求まっている $C_{\ell,R}$ の推定点を仮の値として用いて仮の $A_\ell(t)$ を算出する。時間領域での相関は、既に推定が完了している $A_\ell(t-2s)$ から $A_\ell(t-s)$、(s : セグメント長) において各チャンネルの時間領域での平均値と $A_m(t-s)$ から $A_m(t)$ までのスペクトル形状に対して行う。周波数領域での相関は必要とする周波数領域でのスペクトル形状の知識を知識制御部に対して要求し、受け取った周波数領域でのスペクトル形状と $A_\ell(t)$ により算出した周波数領域でのスペクトル形状により行う。時間領域と周波数領域の 2 つの相関を用いて目的音として妥当と思われる点を推定点として採用する。

2.5.2 音声認識処理

目的音を示す記号 (列) を入力として受け取った知識制御部は、その記号をもとにあらかじめ格納されている目的音の周波数領域でのスペクトル形状の集合を知識群として選択する。知識制御部は波形分離部から要求された知識を波形分離部に送り出す (図 2.4)。

波形分離部では、分離された波形と知識の時刻をどれだけ進めるのが妥当であるのかを決定するために時刻 t において知識内の時刻 $t'+s$ の知識により波形分離を行ったときの周波数領域のスペクトル形状の相関値と同様にの時刻 $t+s$ において知識内の時刻 t' の知識を用いた場合、時刻 t において知識内の時刻 $t'+s$ の知識を用いた場合の相関の中から最も高い値を示す場合をこの時点で尤も目的音らしいと判断しそれぞれの時刻を進めていく (図 2.5)。

認識部では、F0 推定部で推定された F0 の値が話声の F0 として妥当か、そして、波形分離の過程を監視し、分離の過程が妥当であるかどうかを判断する。さらに、分離された波形の周波数領域での形状と知識との間で相関をとり、その平均値で分離結果の妥当性を判断する。最後に、認識部は認識の結果を記号 (列) にて出力する。

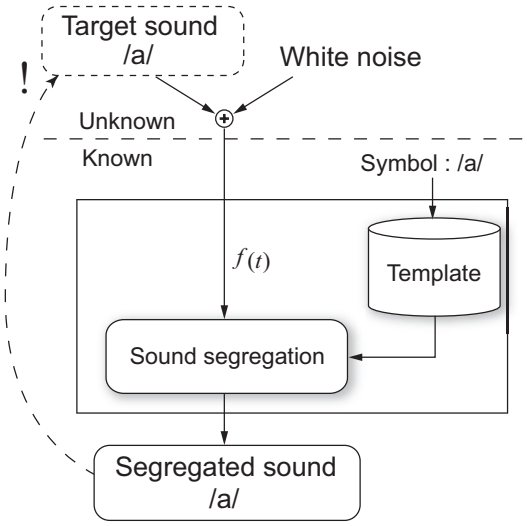
2.6 他の音声認識手法との比較

図 2.6 に本研究の手法と既存の音声認識手法を模式化したものを示す。入力音に対して分析を行ったものと、与えられた知識を用いて、入力音の検定を行い、その検定の結果が認識結果となるという大きな枠組みは、提案手法も既存の手法も変わりがない。これは、未知の言語モデルから生成された音響パラメータ列を観測し、その観測結果と知識を用いて未知の言語モデルを推定するという大きな枠組みとしてはいずれの手法も同じであるからである。しかし、既存の手法と提案手法とには大きな相違点がある。既存の手法は、分析を行った入力音と知識とを比較するだけで認識結果を得ているが、提案手法ではそれに加えて検定の過

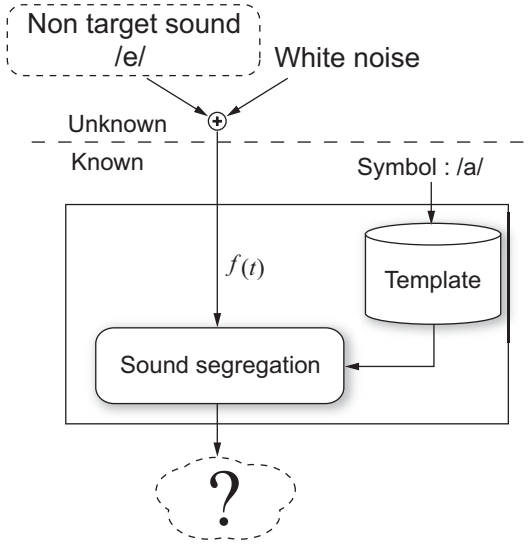
程にも知識を用いている。このように結果にだけでなく、認識の過程にも知識を導入することにより妥当な認識結果を得ようとしている点が、他の音声認識手法とは異なる点である。

音源分離を前処理として用いる音声認識手法と音源分離を認識規範として用いる本研究の手法との比較をする (図 2.6(b) と図 2.6(d))。前処理の目的は認識の入力段階をクリーンな環境に可能な限り近づけることにある。よって、音源分離の性能がシステム全体の認識率に影響を与える事となり、システム全体としては雑音に対応する事が可能になったが認識器そのものが雑音に対してロバストになった訳ではない。また、音源分離を前処理として用いたものは、音源分離と認識が独立した状態になっている。よって、仮に認識器が用いる知識を前処理である音源分離が共有できたとしても、それが認識器そのものの性能に影響を与える訳ではない。一方、本研究の手法は音源分離そのものを認識の枠組みの中に取り入れ規範として用いる。この場合、音源分離の過程において積極的に目的音を探索させ、目的音が存在することが妥当であるかどうかを判断する。認識器そのものに音源分離が取り込まれた事、妥当な過程と結果を求める事から、認識器そのものが雑音に対してロバストになる事が期待できる。

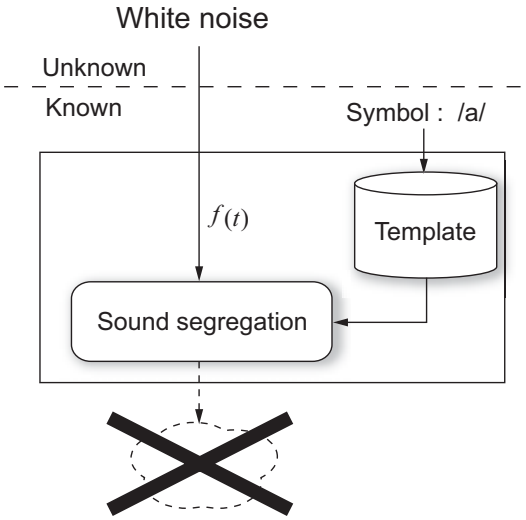
音響モデルを変形する音声認識手法と音源分離を認識規範として用いる本研究の手法との比較をする (図 2.6(c) と図 2.6(d))。音響モデルを変形する音声認識手法では、クリーンな環境に対応した音響モデルを雑音環境に対応させるために音響モデルそのもの変形させる。雑音が既知である場合には、その雑音に適した音響モデルを作成でき、雑音環境が定常であるならば認識率を向上させることができる。しかし、雑音環境が変化して、音響モデルの変形がその状況に適さなくなってしまうと認識率は低下してしまう。これは、システムそのものが音響モデルを周囲の環境に適するように変形するようになっていないことによる。また、音響モデルの変形には雑音 HMM を音響モデルに重畳させる手法が多くとられるが、HMM を用いているため雑音に関して学習を行う必要がある、このため、突発的な雑音や未知の雑音が存在する状況に対応することは困難である。一方、本研究の手法は、周囲の環境が変化したとしても、音源分離処理が積極的に目的音を分離しようと試み、認識を行っている時点で尤もらしいものを分離しようとする。このため、周囲の環境変化に柔軟に対応することが可能で、突発雑音や非定常雑音、未知の雑音が存在する状況でも認識を行えると期待できる。



(a) 入力音中に目的音が存在する場合 (目的音が分離される)



(b) 入力音中に目的音が存在しない場合 (目的音以外の音が分離される)



(c) 入力音中に音声が存在しない場合 (何も分離されない)

図 2.1 音源分離と認識の関係

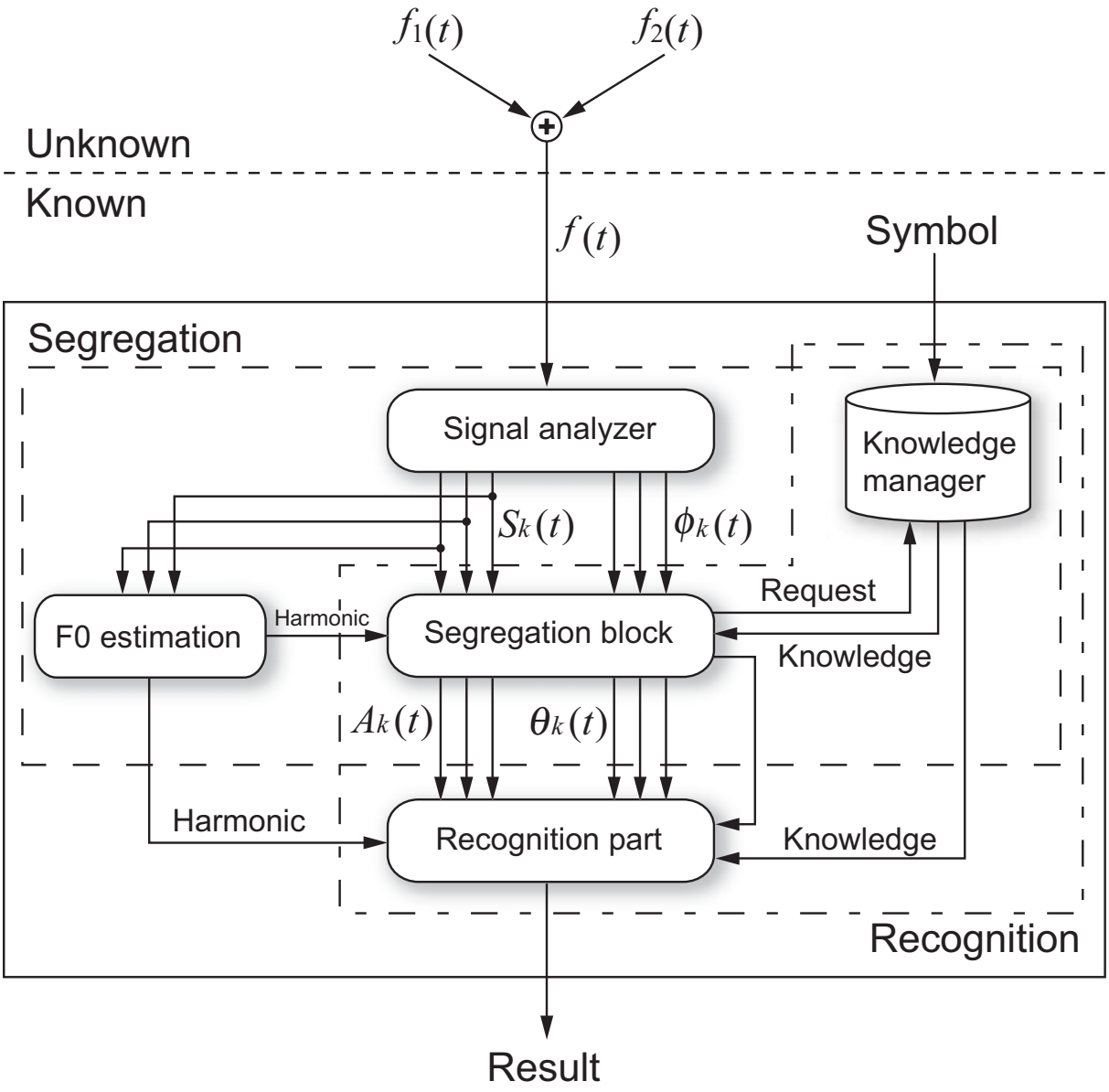


図 2.2 本研究のモデル概要

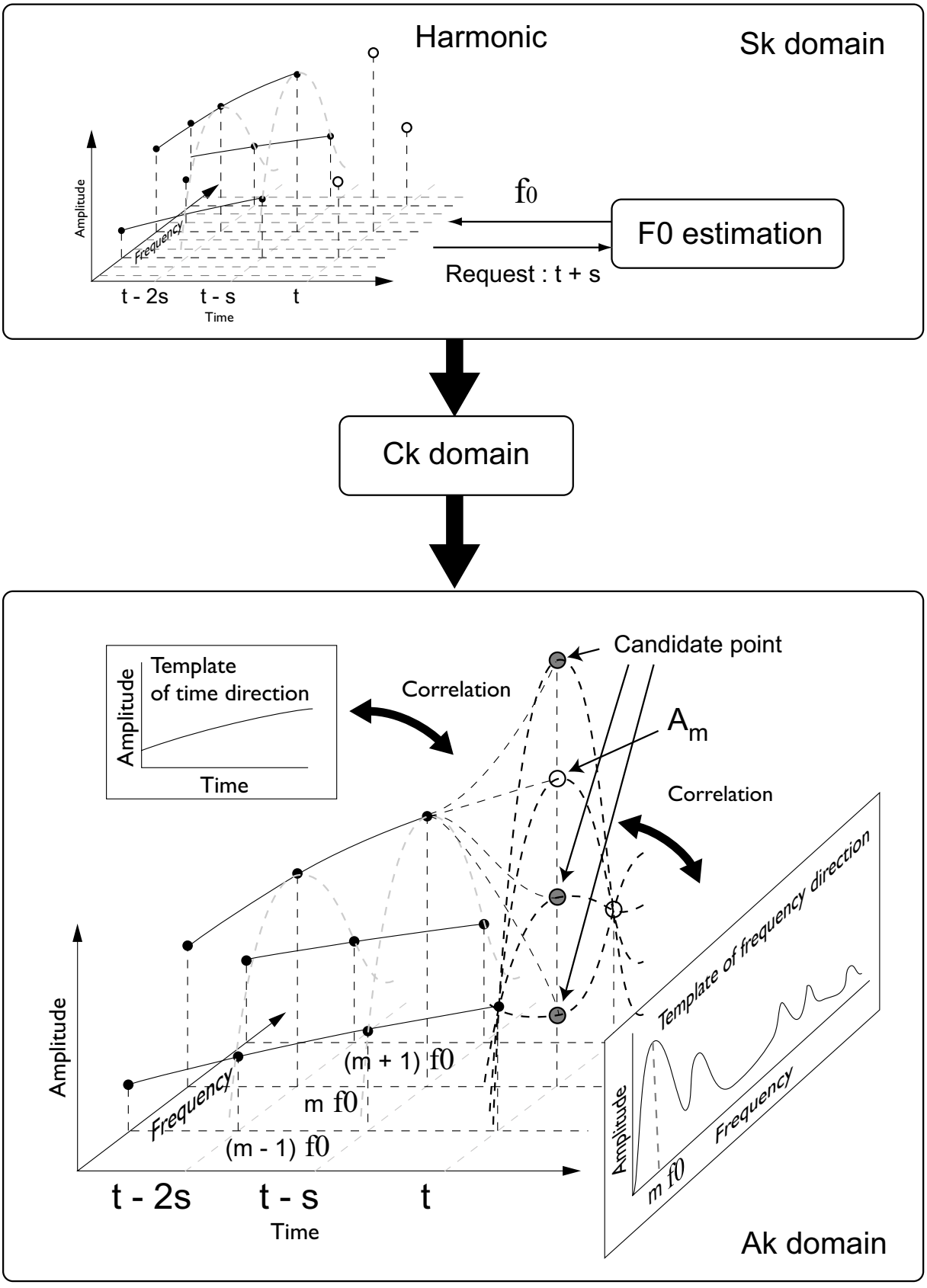

Harmonic
Sk domain
Amplitude
Frequency
t - 2s
t - s
t
Time
f0
F0 estimation
Request : t + s
Ck domain
Template
of time direction
Amplitude
Time
Correlation
Candidate point
Am
Correlation
Template of frequency direction
Amplitude
Frequency
m f0
Amplitude
Frequency
(m + 1) f0
m f0
(m - 1) f0
t - 2s
t - s
t
Time
Ak domain

図 2.3 推定点の決定

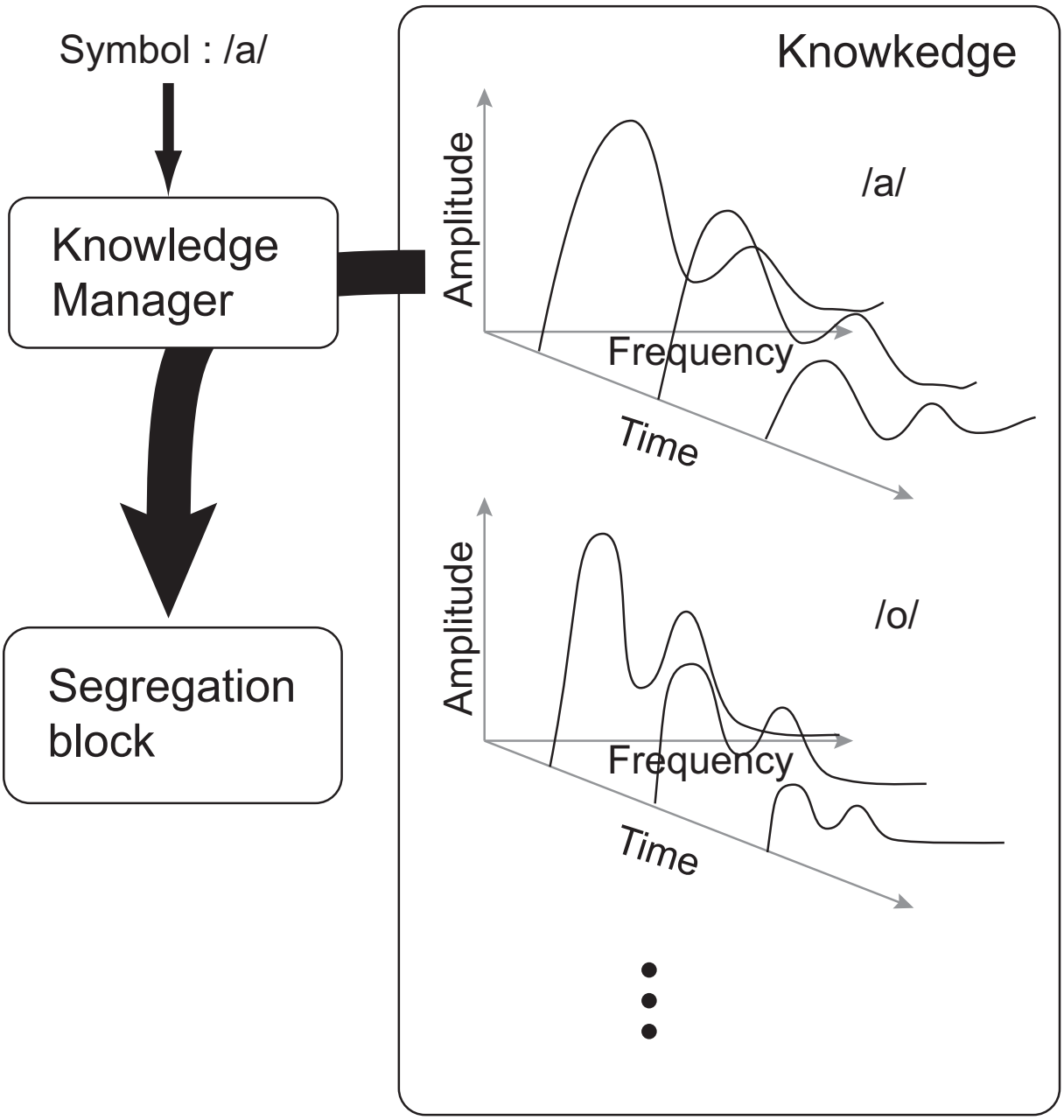


図 2.4 知識制御部の概要

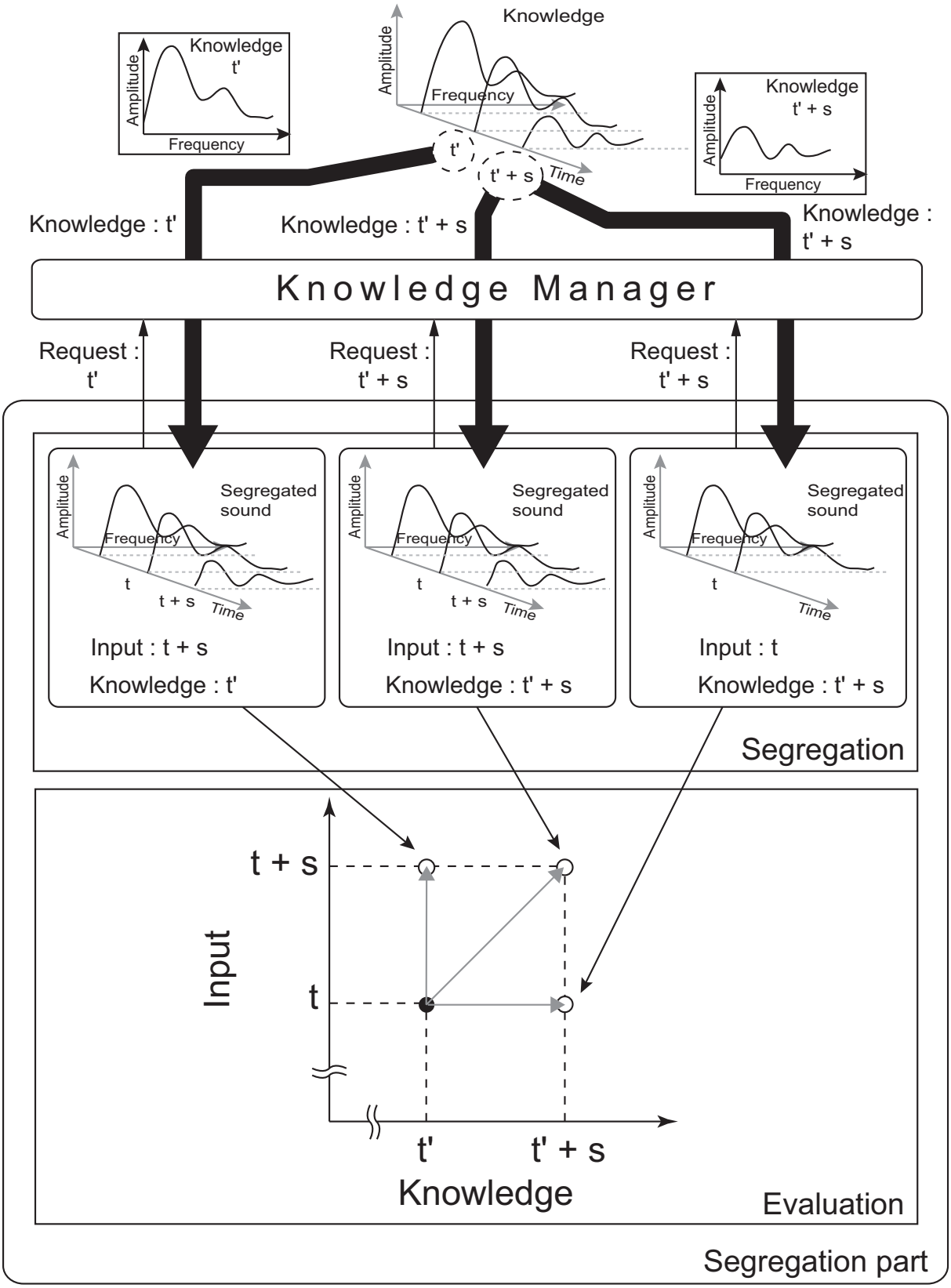
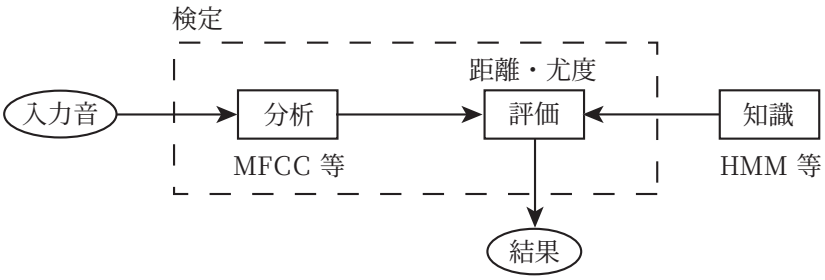
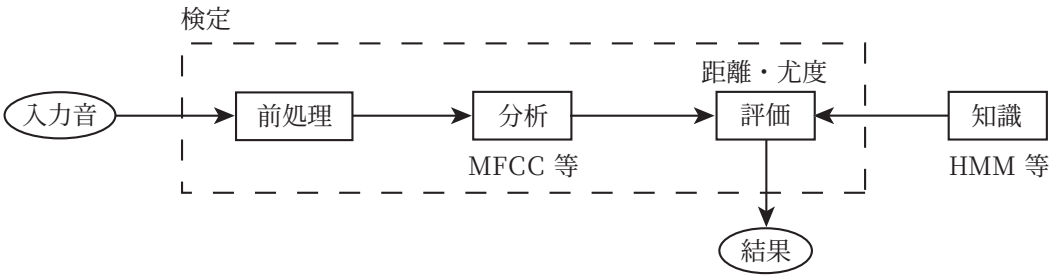


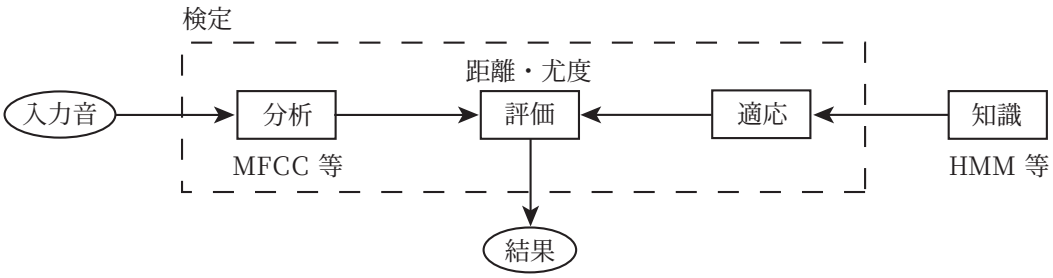
図 2.5 妥当な分離方向の探索



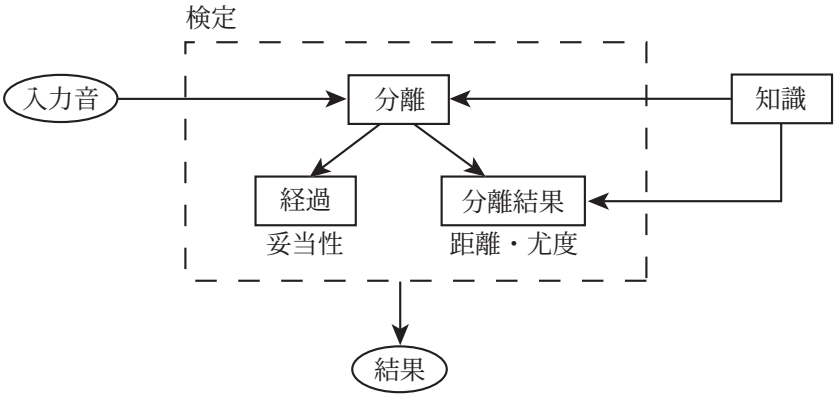
(a) クリーンな環境での認識



(b) 雑音環境下での前処理による認識



(c) 雑音環境下でのモデルの変形による認識



(d) 提案モデル

図 2.6 既存の手法と提案手法との比較

第 3 章　モデルの実装

本章では、前章で示した基本原理に基づいたモデルの実装に関して説明する。

3.1　音源分離処理

3.1.1　信号解析部

信号解析部では入力された信号から瞬時振幅と瞬時位相を抽出し、時間と周波数領域の表現に変換する。本研究での分析の対象は音声である。女性のような高い周波数領域まで調波成分の情報が必要になる事を想定して、解析可能な周波数範囲を 0〜10 kHz とする定帯域フィルタバンクを用いて事とした。よって、信号解析部はサンプリング周波数 20 kHz のデータを入力として受け付ける。さらに、音声の調波成分を扱う事から、各調波成分の重複をさけるためにより狭い周波数帯域で分割されている必要がある。そこで、各フィルタの帯域幅を 20 Hz とした。

表 3.1 にフィルタバンクの仕様を示す。次に、このフィルタバンクが音声を問題なく処理

表 3.1　フィルタバンク設計仕様

フィルタ数	500
解析周波数範囲	0〜10000 Hz
帯域幅	20 Hz
サンプリング周波数	20 kHz

できるかどうかを確認するために、設計したフィルタバンクで音声を分析したものを分析の逆操作による合成フィルタにより再合成して、その結果を比較した。図 3.1 に ATR 音声データベース mau の/a/を用いて分析、再合成した結果を示す。図 3.1(a) は入力音の時間波形である。この音声を分析した結果が図 3.1(b) である。これは、瞬時振幅 $S_k(t)$ を時間―周

波数平面で表現したもので、振幅の大きさをグレースケールで表現しており、黒いものほど振幅が大きい事を示している。次に、この分析結果を分析に用いたフィルタの逆操作を行う合成フィルタにより再合成した信号が図 3.1(c) である。原信号とこの再合成音の差を雑音と見なし SNR を算出したところ、19.73 dB であった。このことから、信号解析部の処理は周波数解析としてはほぼ問題なく行えると考えられる。

3.1.2 F0 推定部

人間の音声のうち母音は基本周波数 (F0) を持っており、その音声は F0 の整数倍の成分で構成されている。したがって、混合音から母音を分離する際には F0 は非常に重要な情報である。本研究の基礎となる窪らのモデルで用いられている F0 推定は持続時間中ほぼ一定であるという仮定の下で推定が行われている。しかし、人間の音声は楽器音と異なり発話時間中、常に変化している。本研究ではバッチ処理ではなくセグメント処理を想定している事から、常に変動している F0 を一定と見なす事は困難である。そこで、今回は定帯域フィルタバンクからの各時刻における出力を周波数領域において自己相関をとる事により F0 を推定する、自己相関法を用いる事とした。これにより、各時刻において F0 を推定する事が可能になり、各時刻での F0 の変化を観測できる。

3.1.3 波形分離部

波形分離部の処理は、鵜木らの二波形分離モデルを用いる。まず始めに、鵜木らのモデルに従い、入力音は目的音と背景音が空間的に加算されたものを仮定する。このとき、入力音を $f(t)$、目的音を $f_1(t)$、背景音を $f_2(t)$ とすると、

$$f(t) = f_1(t) + f_2(t) \tag{3.1}$$

となる。この入力音を K 個の分析フィルタ群を用いて時間と周波数領域の表現に変換する。ここで、k 番目の分析フィルタを通過した $f_1(t)$、$f_2(t)$ を

$$X_{1,k}(t) = A_k(t)\exp(jw_k t + jq_{1k}(t)) \tag{3.2}$$
$$X_{2,k}(t) = B_k(t)\exp(jw_k t + jq_{2k}(t)) \tag{3.3}$$

と、表せるとすると仮定すると入力信号 $f(t)$ が k 番目の分析フィルタを通過すると、

$$X_k(t) = X_{1,k}(t) + X_{2,k}(t) \tag{3.4}$$
$$= S_k(t)\exp(jw_k t + jf_k(t)) \tag{3.5}$$

となる。ただし、w_k は分析フィルタの中心周波数、$A_k(t)$、$B_k(t)$、$S_k(t)$ は瞬時振幅、$f_k(t)$ は瞬時出力位相、q_{1k}、q_{2k} は瞬時入力位相である。また、$S_k(t)$、$f_k(t)$ は、それぞれ、式 (3.2)〜(3.5) から

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)\cos q_k(t) + B_k^2(t)} \tag{3.6}$$

$$f_k(t) = \arctan\left(\frac{A_k(t)\sin q_{1k}(t) + B_k(t)\sin q_{2k}(t)}{A_k(t)\cos q_{1k}(t) + B_k(t)\cos q_{2k}(t)}\right) \tag{3.7}$$

である。よって、$A_k(t)$、$B_k(t)$ は、それぞれ

$$A_k(t) = \frac{S_k(t)\sin(q_{2k}(t) - f_k(t))}{\sin q_k(t)} \tag{3.8}$$

$$B_k(t) = \frac{S_k(t)\sin(f_k(t) - q_{1k}(t))}{\sin q_k(t)} \tag{3.9}$$

となる。ここで、$q_k(t) = q_{2k}(t) - q_{1k}(t)$ であり、$q_k(t) \neq np, n \in \mathbf{Z}$ とする

ここで、式 (3.8)、(3.9) をみると、観測により既知であるパラメータが $S_k(t)$ と $f_k(t)$ であり、未知のパラメータが $A_k(t)$、$B_k(t)$、q_{1k}、q_{2k} の 4 つであるのに対して方程式は 2 つしかない。したがって、すべてのパラメータを同時にしかも一意に決定することは不可能である。1.3 節 でも触れたように、与えられた情報量に対して欠落した情報量が多い状態となっている二波形分離問題は不良設定の逆問題であることにこれは起因する。そこで、鵜木らの二波形分離モデルでは、表 3.2 に示した Bregman が提唱している聴覚情景解析に基づく 4 つの発見的規則 [Bre90, Bre93] をこの問題の制約条件としている。これらの定性的規則は、以下のような定式化を行って実際は運用している。

[制約条件 1] (立ち上がり · 立ち下がりの同期)

F0 の立ち上がり時刻を T_S、立ち下がり時刻を T_E とする。このとき、同じ音源で生じた信号成分であれば、k 番目の高調波の立ち上がり時刻 $T_{k,\mathrm{on}}$ と立ち下がり時刻 $T_{k,\mathrm{off}}$ は F0 の立ち上がり、立ち下がり時刻と一定の範囲内で一致していなくてならないと考えられるので、

$$|T_S - T_{k,\mathrm{on}}| \leqq \Delta T_S \tag{3.10}$$

$$|T_E - T_{k,\mathrm{off}}| \leqq \Delta T_E \tag{3.11}$$

となるべきである。

[制約条件 2] (漸近的変化 (多項式近似))

ある区間における瞬時振幅 $A_k(t)$、瞬時入力位相 $q_{1k}(t)$、基本周波数 $F_0(t)$ のそれぞれの導

関数が

$$\frac{dA_k(t)}{dt} = C_{k,R}(t) \tag{3.12}$$

$$\frac{dq_{1k}(t)}{dt} = D_{k,R}(t) \tag{3.13}$$

$$\frac{dF_0(t)}{dt} = E_{0,R}(t) \tag{3.14}$$

と表されるものとする。ただし、$C_{k,R}(t)$、$D_{k,R}(t)$、$E_{0,R}(t)$ は、区分的に微分可能な R 次多項式である。このとき、$A_k(t)$、$q_k(t)$、$F_0(t)$ は、それぞれ、

$$A_k(t) = \int C_{k,R}(t)dt + C'_{k,0} \tag{3.15}$$

$$q_{1k}(t) = \int D_{k,R}(t)dt + D'_{k,0} \tag{3.16}$$

$$F_0(t) = \int E_{0,R}(t)dt + E'_{0,0} \tag{3.17}$$

と表せる。ただし、$C'_{k,0}$、$D'_{k,0}$、$E'_{0,0}$ は積分定数である。

[制約条件 3] (漸近的変化 (なめらかさ)

閉区間 $[t_a, t_b]$ における $A_k(t)$ と $q_{1k}(t)$ に対して、定積分

$$s_A = \int_{t_a}^{t_b} \left[A_k^{(R+1)}(t)\right]^2 dt \tag{3.18}$$

$$s_q = \int_{t_a}^{t_b} \left[q_{1k}^{(R+1)}(t)\right]^2 dt \tag{3.19}$$

が、最小になるとき、$A_k(t)$、$q_{1k}(t)$ を最も滑らかであるとする、ただし、$A_k(t)$ と $q_{1k}(t)$ は、それぞれ、式 (3.12) の $C_{k,R}(t)$ と式 (3.13) の $D_{k,R}(t)$ を用いて決定された瞬時振幅と瞬時位相である。また、$A_k^{(R+1)}(t)$ と $q_k^{(R+1)}(t)$ は、それぞれ、$A_k(t)$ と $q_{1k}(t)$ の $(R+1)$ 次の導関数である。

[制約条件 4] (調波関係)

F0 を $F_0(t)$、高調波の次数を N_{F_0} とする。このとき、調波関係にある信号は

$$n \times F_0(t), \qquad n = 1, 2, \ldots, N_{F_0} \tag{3.20}$$

の関係を満たさなければならない。

[制約条件 5] (時間領域での振幅包絡 $A_k(t)$ 間の相関)

s はセグメントの長さ、ℓ は ℓ 次高調波のチャンネルを表すものとし、$\ell = 1, 2, \ldots, N_{F_0}$ とする。このとき、時刻 $t-s$ から t までの振幅包絡 $A_m(t)$、$(m \in \ell)$ と時刻 $t-2s$ から $t-s$ ま

での振幅包絡 $A_\ell(t)$ の時間領域での平均は、その形状が類似しなければならない。

$$A_{\text{mean}}(n) = \sum_\ell A_\ell(n)/N_{F_0}, \qquad t-2s \leqq n \leqq t-s \tag{3.21}$$

$$\frac{A_{\text{mean}}(n)}{\|A_{\text{mean}}(n)\|} \approx \frac{A_m(n')}{\|A_m(n')\|}, \qquad t-s \leqq n' \leqq t \tag{3.22}$$

ただし、$\|\cdot\|$ はノルム記号である。

本研究ではこれらに加えて知識と周波数領域の相関を考慮する。

ℓ は調波位置のチャンネルを表し、$\ell = 1,2,\ldots,N_{F_0}$ とする。ある時刻 t における $A_\ell(t)$ の周波数領域での包絡 $Af(t)$ と知識上の時刻 t' における周波数領域での包絡 $Af_{\text{template}}(t')$ は、その形状が類似しなければならない。

$$\frac{Af(t)}{\|Af(t)\|} \approx \frac{Af_{\text{template}}(t')}{\|Af_{\text{template}}(t')\|} \tag{3.23}$$

本論文では、人間が位相の変化に対して敏感でなく分離に関して重要な位置を占めないとう立場と SNR の高い分離音を抽出する事を目的としないという立場から位相の時間変化はすべて背景音において生じるものと仮定し、式 (3.13) における $D_{k,R} = 0$ とした。以上を用いて、波形分離部では次のような手順で目的音分離を行う。

1. F0 推定部において F0 が推定されたセグメントに対して信号の立ち上がり時刻 T_S、立ち下がり時刻 T_E を求める。ただし、F0 が推定されるが、セグメントが信号の立ち上がりもしくは立ち下がり部分にない場合は、T_S はセグメントの開始点、T_E はセグメント終了点とする。
2. 基本周波数 $F_0(t)$ の調波関係を満たす $X_k(t)$ を求める。
3. Kalman filter を用いて、式 (3.12) の $C_{k,0}$ と誤差 $P_k(t)$ を推定する。
4. 誤差範囲内 $\hat{C}_{k,0} - P_k(t) \leqq C_{k,1}(t) \leqq \hat{C}_{k,0} + P_k(t)$ から、時間方向に spline 補間された $C_{k,1}$ の候補を求める。
5. $C_{k,1}$ の候補を用いて対応する $A_k(t)$ の候補を求める。
6. 時間領域においては $A_k(t)$ を現在のセグメントと 1 つ前の区間内で spline 補間を行い、時間領域での振幅包絡を求める。
7. 周波数領域においては時刻 t における $A_k(t)$ の調波位置の値を用いて spline 補間を行い、周波数領域での振幅包絡を求める。
8. 6 と 7 のそれぞれに対して、式 (3.22) と式 (3.23) による相関を求める (図 2.3)。

9. 8 の時間領域の相関と周波数領域の相関の和を候補点選択の指標として設定し、この指標が最大となる点を候補点とする。
10. 以上の処理から最終的な $A_k(t)$ を求める。
11. 周波数領域において 4〜10 を繰り返す。
12. 時間領域において 1〜11 を繰り返す。

3.2 音声認識処理

3.2.1 知識制御部

知識制御部は、目的音を表す音素表記を入力として受け取る。本研究で扱う “知識” は音素表記そのものやその他の記号列からなるものではなく、音素表記に対応する音声の振幅包絡の集合の事を指していて、音声に関する物理量を知識として扱っている。そこで、音素表記のような受け取った記号列をもとに知識制御部は、あらかじめ格納されている目的音の周波数領域での振幅包絡の集合を知識群として選択する (図 2.4)。

知識は、あらかじめクリーンな音声をはじめに周波数解析部で用いているものと同じ定帯域フィルタバンクに通し、瞬時振幅と瞬時位相を抽出し時間と周波数領域の表現に変換する。抽出した瞬時振幅を F0 推定部と同様の F0 推定を行い、調波関係の成分を分離する。各セグメント位置において調波関係の瞬時振幅を周波数領域で spline 補間を行う。すべての時刻において、この操作を繰り返す。各セグメント位置における振幅包絡は知識作成時の時刻によりラベル付けされ、知識内での時刻を呼び出す事により自由な位置の知識を呼び出す事ができるようになっている。この振幅包絡群を知識群として知識制御部は格納している。

知識制御部は、波形分離部から知識内の時刻の形で要求され、該当する知識を波形分離部に対して提供する。

3.2.2 波形分離部

波形分離部では、音源分離部の処理に従って目的音の分離を行う。このとき、波形分離部では目的音としてより尤もらしいものを分離するために、入力信号の時刻 t と知識内での時刻 t' の組み合わせをかえながら周波数領域での相関をとりながら分離を試行し、その時点で目的音として尤もらしい分離音を目的音として、次のステップでの分離に移る。この判断は、ある時刻 t で音源分離処理において、調波位置の $A_k(t)$ の推定が終了した後に行われる。セグメントの長さを s としたとき、1 回の分離で試行する信号と知識の組は、入力信号の時刻 t と知識内での時刻 t'、$t+s$ と t'、t と $t'+s$ である (図 2.5)。

この 3 つの中からいずれの方向に進んでいったのか表記する。このような表記に従って、最終的に波形分離をおこなた例を図 3.2 に示す。この図は、動的計画法における時間伸縮関数や HMM におけるトレリスのような表現と告示している。このことから、波形分離のときに 3 つの組み合わせのうちいずれが目的音として尤もらしい方向であるかを決定するには、2 つの波形間の距離を最小にするように進んでいく方策をとるのであれば動的計画法、3 つの状態の遷移に状態遷移確率を用いればビタービアルゴリズムを用いる事により、分離過程がより妥当に行われているかを議論できる。

そこで、本論文では 2 つの波形間の距離を細小にするように進んでいく方策をとり動的時間伸縮 (Dynamaic Time Wraping; DTW) を用いて、より妥当な分離を行うようにする。本論文では、横軸を知識における時刻 t'、縦軸を信号における時刻 t とするような時間伸縮関数の表現を用いる。u 回目 $(1 < u \leqq U)$ の試行における信号の時刻 $t(1 \leqq t \leqq T)$ と知識内での時刻 $t'(1 \leqq t' \leqq T')$ において、2 つの振幅包絡の周波数領域での正規化した相互相関の最大値を $R_a(u) = R(t', t)$、$(a = 1, 2, 3)$ とする。ここで、s をセグメント長さとして、

$$R_1(u) = R(t' + s, t + s) \tag{3.24}$$
$$R_2(u) = R(t', t + s) \tag{3.25}$$
$$R_3(u) = R(t' + s, t) \tag{3.26}$$

と定義する。このとき、2 つの信号間の距離が最も近くなるのは、相関値が最も高いものであるので、尤も妥当な方向の相互相関の最大値 $R_{\max}(u)$ を

$$R_{\max}(u) = \underset{a}{\operatorname{argmax}} \left(R_a(u) w(u) \right) \qquad , a = 1, 2, 3 \tag{3.27}$$

と定義する。このときの a をパス番号とする。パス番号 1 は斜めに右上に進み、パス番号 2 は上方に、パス番号 3 は右方へ進む事を表す。ここで、$w(u)$ は u 回めの試行に関する重みで、

$$w(u) = \begin{cases} 1, & \text{信号の時刻 t において F0 が推定されている} \\ 0, & \text{信号の時刻 t において F0 が推定されなかった} \end{cases} \tag{3.28}$$

となっている。

このような、時間伸縮関数に対して本論文では以下のような制約を設けた。

1. 始終端の制約
2. 単調性の制約
3. 傾斜制限

試行 u 回目の信号の時刻を $j_t(u)$ 知識内の時刻を $j_{t'}(u)$ とすると。

$$始点: j_t(1) = 1, j_{t'}(1) = 1 \tag{3.29}$$

$$終点: j_t(U) = T, j_{t'}(U) = T' \tag{3.30}$$

が、始終端の制約になる。単調性の制約は、時間伸縮は正の方向のみにしか起きず、負の方向の時間伸縮は起きないというものである。

$$j_t(u+1) \geqq j_t(u) \tag{3.31}$$

$$j_{t'}(u+1) \geqq j_{t'}(u) \tag{3.32}$$

傾斜制限はパスが極端な時間伸縮が起きないように制限するもので、本論文ではパスが連続して 3 回パス番号 2 または 3 の方向に進む事を制限している。始終端の制約と単調性の制約はシステムが強制的に制約を満たすように分離を進めていく。傾斜制限は制限の条件に達した場合パスの方向を強制的に変更する。

3.3 認識部

認識部では、F0 推定部で推定された F0 の値が話声の F0 として妥当か、そして、波形分離の過程を監視し、分離の過程が妥当であるかどうかを判断する。さらに、分離された波形の周波数領域での形状と知識との間で相関をとり、その平均値で分離結果の妥当性を判断する。

F0 の妥当性は、人間の通常の発話時の F0 は 60〜400 Hz 程度であるという立場に立ち、その範囲を大きく超える基本周波数が推定された場合には人間の音声ではないと判定する。波形分離部での過程の妥当性は迫江と千葉により提唱された全体的なパスの制約により判断する。これは

$$|j_t(u) - j_{t'}| \leqq T_0 \tag{3.33}$$

と表現され、図 3.3 のグレーの領域以外をパスが異常な方向へ進んでいると判断する。この範囲を超えて分離が進んだ場合は、分離が妥当に行われなかったとして、認識部は目的音が存在しないと判断し処理を終了する。

分離結果の妥当性は、分離によって得られた分離音と知識との全体的な距離により評価す

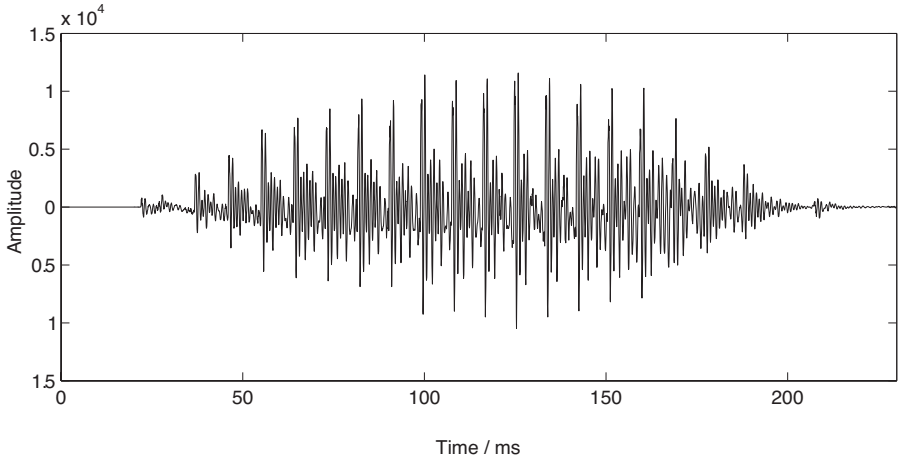
る。具体的には、音源分離部で用いた相互相関の最大値を用いて以下のように算出する。

$$R_{\text{mean}} = \sum_{u=1}^{U} R_{\max}(u) / \sum_{u=1}^{U} w(u) \tag{3.34}$$

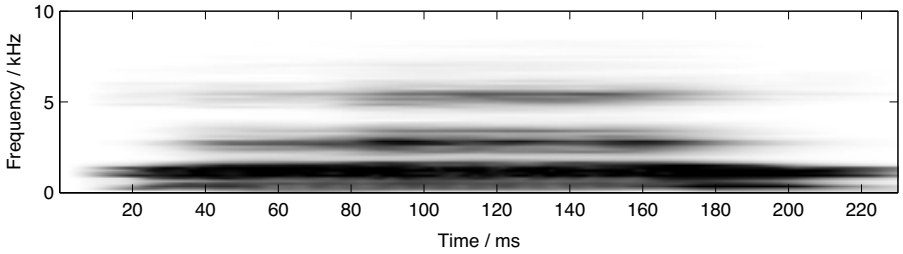
$$= \sum_{u=1}^{U} \underset{a}{\text{argmax}} \left(R_a(u) w(u)\right) / \sum_{u=1}^{U} w(u) \qquad , a = 1,2,3 \tag{3.35}$$

R_{mean} がしきい値 ($R_{\text{threshold}}$) を下回った場合は、分離結果が妥当ではないとして、認識部は目的音が存在しないと判断し処理を終了する。

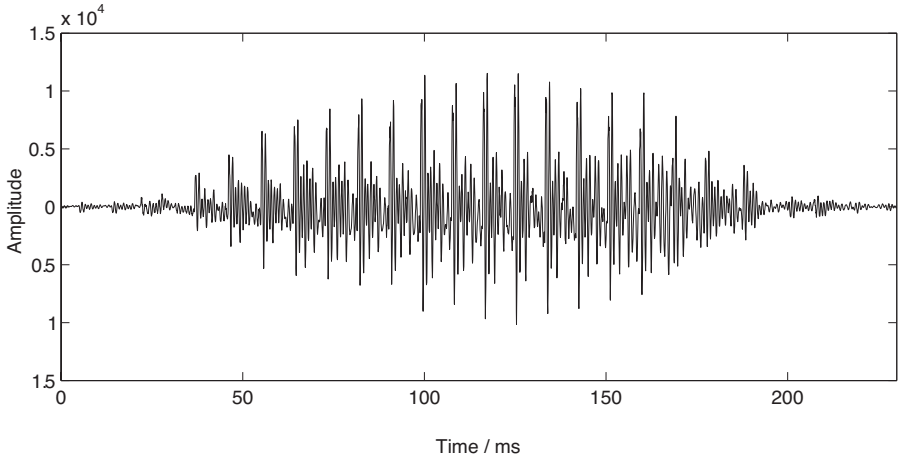
以上のような妥当性に関する制約をクリアした場合に、システムは目的音が存在したとして、その旨を出力する。



(a) 原信号



(b) 瞬時振幅 $S_k(t)$



(c) 再合成信号

図 3.1 フィルタバンクの評価

表 3.2　Bregman の発見的規則と制約条件

発見的規則 (Bregman, 1993) [McA93]	制約条件 (鵜木、赤木、1999) [鵜木 9
i) 関連のない音が一緒に始まったり、終わったりすることはない	立ち上がり立ち下がりの同期
ii) 変化は急激におこらない	漸近的変化
a) 一つの音の属性は、ゆっくりと滑らかに変化する傾向がある	多項式近似なめらかさ
b) 同じ音源から生じる音の一連の音の属性は、ゆっくりと滑らかに変化する傾向にある	多項式近似なめらかさ
iii) 物が繰り返し振動するときには、共通の基本周波数 (F0) の整数倍の音響的成分が発生する	調波関係
iv) 一つの音響事象に生じる多くの変化は、その音を構成する各成分に同じような影響を与える	振幅包絡間の相関

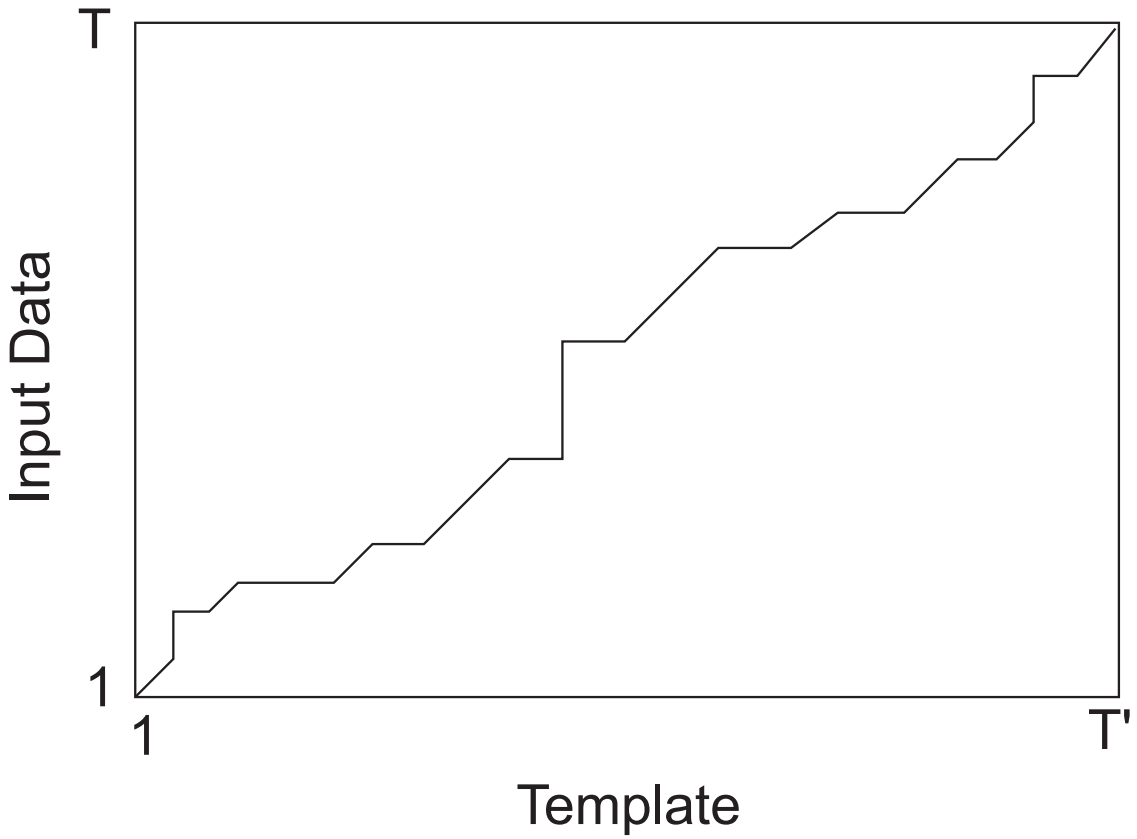


図 3.2 パスの遷移例

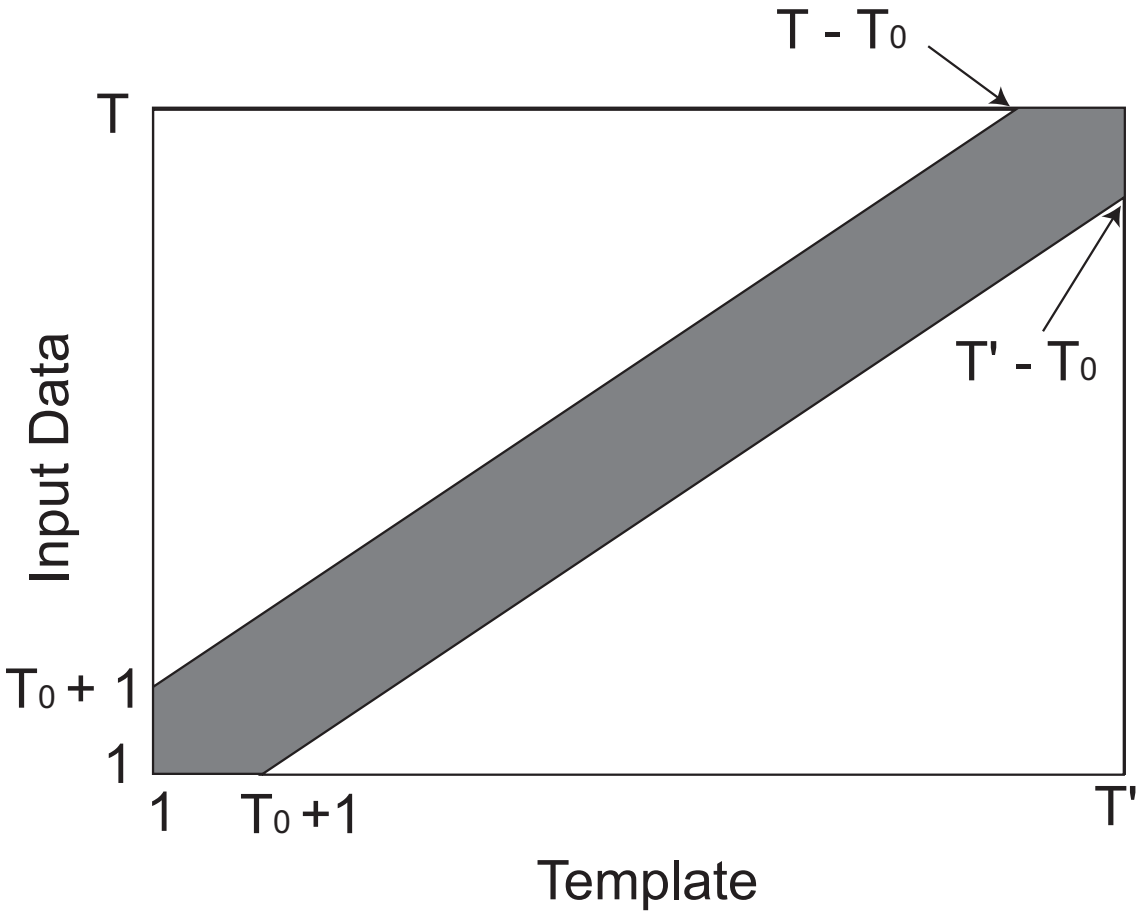


図 3.3 全体なパスの制約

第 4 章　評価

提案手法の評価をおこなうためにシミュレーションを行った。本章では、その条件と結果、有効性の検討を行った結果を示す。

4.1　提案手法の評価

まずはじめに、本研究で提案した音源分離を認識の規範とし妥当な認識結果を得るというコンセプトの有効性を確認するためのシミュレーションを行った。

第 3 章で実装したモデルは、本研究のコンセプトの核となる波形分離部と認識部のほかに、知識制御部、信号解析部、F0 推定部から構成されている。このため、本研究のコンセプトの有効性を確認するためには、モデルを切り分けて考える必要がある。そこで、実装したモデルの中で入力音の影響を受ける F0 推定部をあらかじめ推定されたものを使うように変更して、波形分離部と認識部の有効性を確認する。第 3 章において、提案手法を実装したシステムに対して F0 推定部のみを変更したシステムを新たに構築した。そのモデル図を図 4.1 に示した。第 3 章で実装したモデルとの違いは F0 推定部を F0 データベース部とした点であり、それ以外の部分に変更はない。よって、新たに作成したモデルと第 3 章で実装したモデルとの間には、枠組み上の違いはなく等価なモデルである。

処理の流れの概要を図 4.2 に示した。F0 データベース部では F0 推定を行わず、あらかじめ作成した 5 母音の F0 推定の結果を格納している。格納している F0 の値はあらかじめ F0 推定を行った際の時刻でラベル付けされている。システムが処理する混合音と格納している F0 の時刻は時間歪みがないようにしてある。波形分離部は処理を行っているデータの時刻を渡す形で F0 データベース部に対して要求を行い、推定 F0 の値を受け取る。この推定 F0 の値を用いて、第 3 章で実装したモデルと同様に波形分離部は入力データと知識内の時刻をより尤もらしい方向へ進めようとして分離を行う。

4.1.1 シミュレーション条件

評価を行うために、DTW (Dynamic Time Wrapping) を用いた手法との比較を行った。DTW による手法は、提案モデルで用いている信号解析部と知識制御部を用いて DTW により認識を行う音声認識システムを作成し、それを用いた。図 4.3 に DTW によるシステムのモデル図を示した。このシステムは、提案手法をもとにした音声認識システムと同様に 3.3 節に示した始終端の制約、単調性の制約、傾斜制限の 3 つの制約を設けている。

2 つのシステムを比較するために、背景音として定常な白色雑音が存在する状況で 1 名の話者が日本語単母音を発話している状況を想定し、このような状況で特定話者認識を行った。シミュレーション条件の詳細を以下に示す。

話者は ATR 音声データベースの mau とし、単母音/a/、/i/、/u/、/e/、/o/ を目的音として用いた。雑音は、ガウス性白色雑音を用いた。システムへ入力する混合音は、上記の目的音と雑音を線形加算し、SNR が、20、10、0、-10、-20 dB となるように調整したものを用いた。混合音のデータは 20 kHz サンプリング、16 bit 量子化されたものを用いた。図 4.4 に、クリーンな環境での mau の /a/ の時間波形、ガウス性白色雑音の時間波形、話声と雑音を SNR が 0 dB となるように調整した混合音の時間波形、混合音を波形解析部のフィルタバンクを通すことによって得られる瞬時振幅 $S_k(t)$ を時間—周波数平面で表現したものを示した。瞬時振幅の図は、振幅の大きさをグレースケールで表現しており、黒いものほど振幅が大きいことを示している。

認識に用いる知識は、クリーンな環境での mau の音声を用いて、3.2.1 節で示した方法により作成したものを用いる。いずれの音声認識システムも同じ知識を知識制御部に格納している。。F0 データベース部には、 入力音として用いる混合音のうち SNR が 0 dB の混合音を第 3 章で実装したモデルと同様の自己相関法により推定した値を格納した。

いずれのシステムも入力音に対して 5 母音の知識を用いて認識を行い、その中から尤もらしいものが混合音中に存在するものとし、それを認識結果とした。 3.3 節で示した提案手法をもとにした音声認識システムの認識部が評価に用いるパラメータとして、表 4.1 に示す値を用いた。

表 4.1　認識部で用いたパラメータ

全体のパスの制約 (T_0)	4
相互相関のしきい値 $R_{\text{threshold}}$	0.90

2 つの手法を比較するための尺度として、正解であるべき状況で行った認識に対して式 (4.1) に示すものを用いた。認識の過程における u 回目 $(1 \leqq u \leqq U)$ の試行の結果得られる知識と入力音の相互相関関数の最大値を $R_{\max}(u)$ としたときに、評価尺度を

$$F = \sum_{u=1}^{U} \left(R_{\max}(u) w(u)\right) / \sum_{u=1}^{U} w(u) \tag{4.1}$$

と定義する。ただし、$w(u)$ は u 回目の試行における重みで、提案手法をもとにした音声認識システムでは、

$$w(u) = \begin{cases} 1, & \text{信号の時刻 t において F0 が推定されている} \\ 0, & \text{信号の時刻 t において F0 が推定されなかった} \end{cases} \tag{4.2}$$

であり、DTW による音声認識システムでは、

$$w(u) = 1 \tag{4.3}$$

とした。この評価尺度は、入力音と知識との間の相関値をもとにしたもので、数値が 1 に近いほど認識結果が確からしいことを表す。

4.1.2 シミュレーション結果

シミュレーション結果を示す。いずれのシステムにおいても、0 dB まで 5 母音の中から混合音中の母音を正しく認識する事ができた。さらに、提案手法にもとづくモデルは -10 dB まで正しく認識することができた。その一方で、DTW によるシステムは/a/と/e/のみを認識したが、その他の /i/、/u/、/o/ に関しては誤認識した。図 4.5 は、式 (4.1) に示した評価値をプロットしたものである。これは、値が 1 に近いほど認識結果が確からしいことを表している。提案手法にもとづくモデルは SNR が小さくなっているにもかかわらず評価値がおおむね 0.90 以上の値でほぼ一定の値を示しているのに対して、DTW による認識手法では SNR の減少に伴い、評価値が著しく減少する傾向がある。2 つのシステムにおける評価値の差は 20 dB では 0.1 程度であったが、-10 dB では 0.6 程度と差が大きくなっている。これは、提案手法をもとにした認識システムは SNR が低い状態においても、より確からしい認識結果を導きだしていることを示している。一方、DTW による手法は SNR が低くなるにつれて、認識結果が確からしくないことを示している。このことから、提案手法にもとづくモデルは、DTW による認識システムとは、結果の確からしさの点で大きく上回っていることがわかる。このシミュレーションから -10〜20 dB の間で提案手法をもとにした音声認識システムは、常に確からしい正解を導くことがわかった。以上により、本研究で提案した音

源分離を認識の規範とし妥当な認識結果を得るというコンセプトの有効性を確認することができた。

4.2 実装モデルの評価

前節のシミュレーションで、本研究で提案したコンセプトが有効であることが確認できた。そこで次ぎに、第 3 章で実装したシステムの性能評価を行った。

4.2.1 シミュレーション条件

前節でのシミュレーションと同様の条件で、第 3 章で実装したシステムを雑音環境下で用いた場合をシミュレーションした。認識方法、認識部でのパラメータ、評価尺度も前節のシミュレーションと同じものと同じとした。比較に用いた DTW にもとづくシステムも同様のものとした。前節のシミュレーションとの違いは、提案手法にもとづいたモデルがシステムに入力された音から逐次的に F0 推定を行うという点だけである。

4.2.2 シミュレーション結果

シミュレーションの結果を示す。いずれのシステムにおいても、0 dB まで 5 母音の中から混合音中の母音を正しく認識する事ができた。図 4.6 は、式 (4.1) に示した評価値をプロットしたものである。前節でのシミュレーション同様に、提案手法は SNR が小さくなっているにもかかわらず評価値がおおむね 0.95 以上の値でほぼ一定の値を示しているのに対して、DTW による認識手法では SNR の減少に伴い、評価値が著しく減少する傾向がある。2 つのシステムにおける評価値の差は 20 dB では 0.1 程度であったが、0 dB では 0.4 程度と差が大きくなっている。これは、提案手法をもとにした認識システムでは、音源分離を認識の規範とし妥当な認識結果を得るというコンセプトが有効に働いているからと考えられる。

前節でのシミュレーションでは、-10 dB においても正しく認識していたにも関わらず、-10 dB において提案手法をもとにしたシステムは認識に失敗した。システムの状態を観察すると、この原因は入力信号の全区間において妥当な F0 を推定できなかったことによる。以上のシミュレーションは、本研究のコンセプトを -10 dB においても有効に働かせるためには、現在用いている自己相関法による F0 推定法より雑音に対しロバストな手法を取り入れる必要があることを示している。雑音に対してロバストな F0 推定法は数多く研究されている [Ish01, 阿部 00, 阿部 96]。これらの F0 推定法を取り入れることにより、この問題は解決できると思われる。したがって、雑音に対してロバストな F0 推定法を用いることにより、

本研究の提案手法を有効に活用できるモデルを構築できると考えられる。

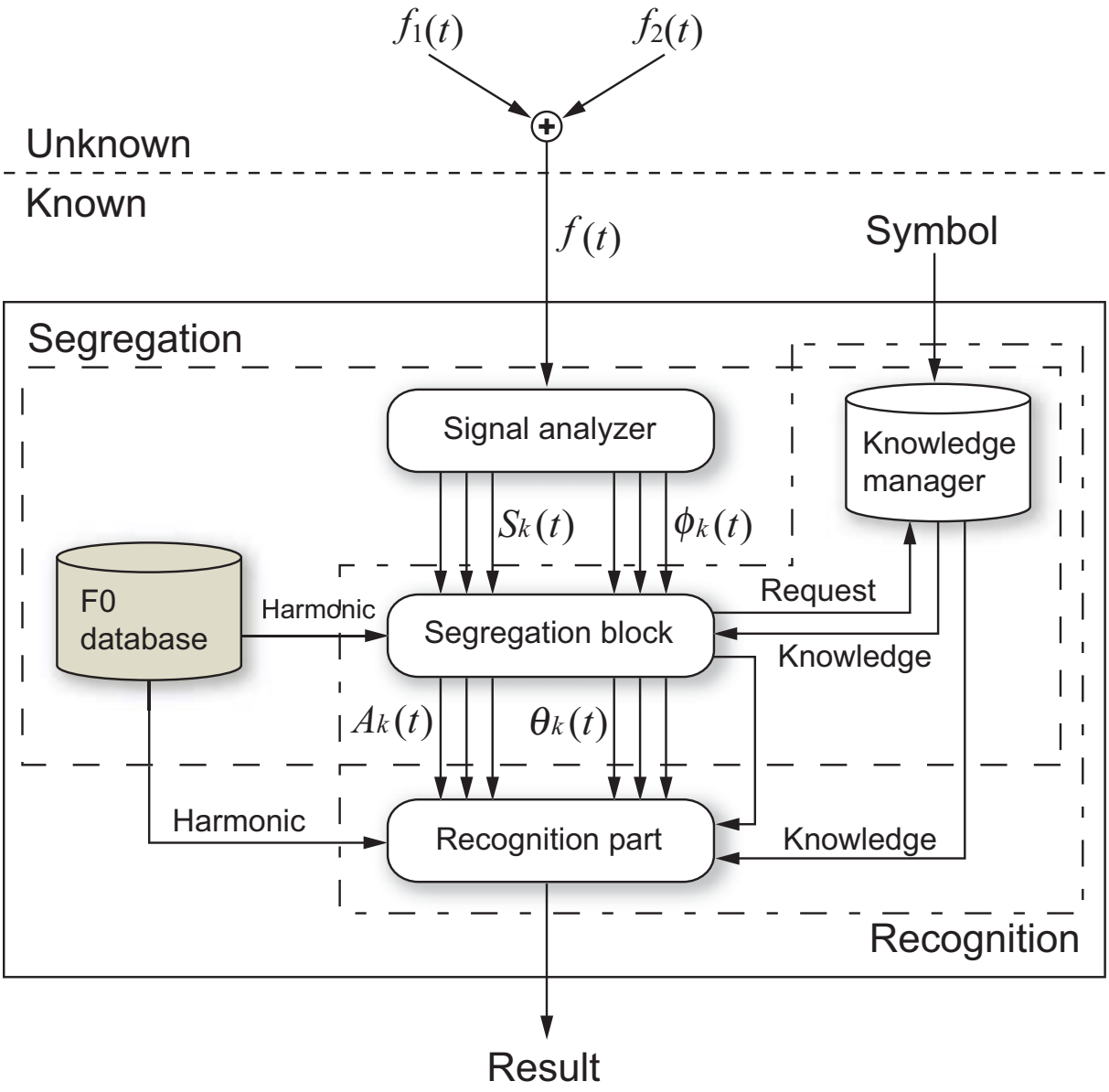


図 4.1 評価用モデルの概要

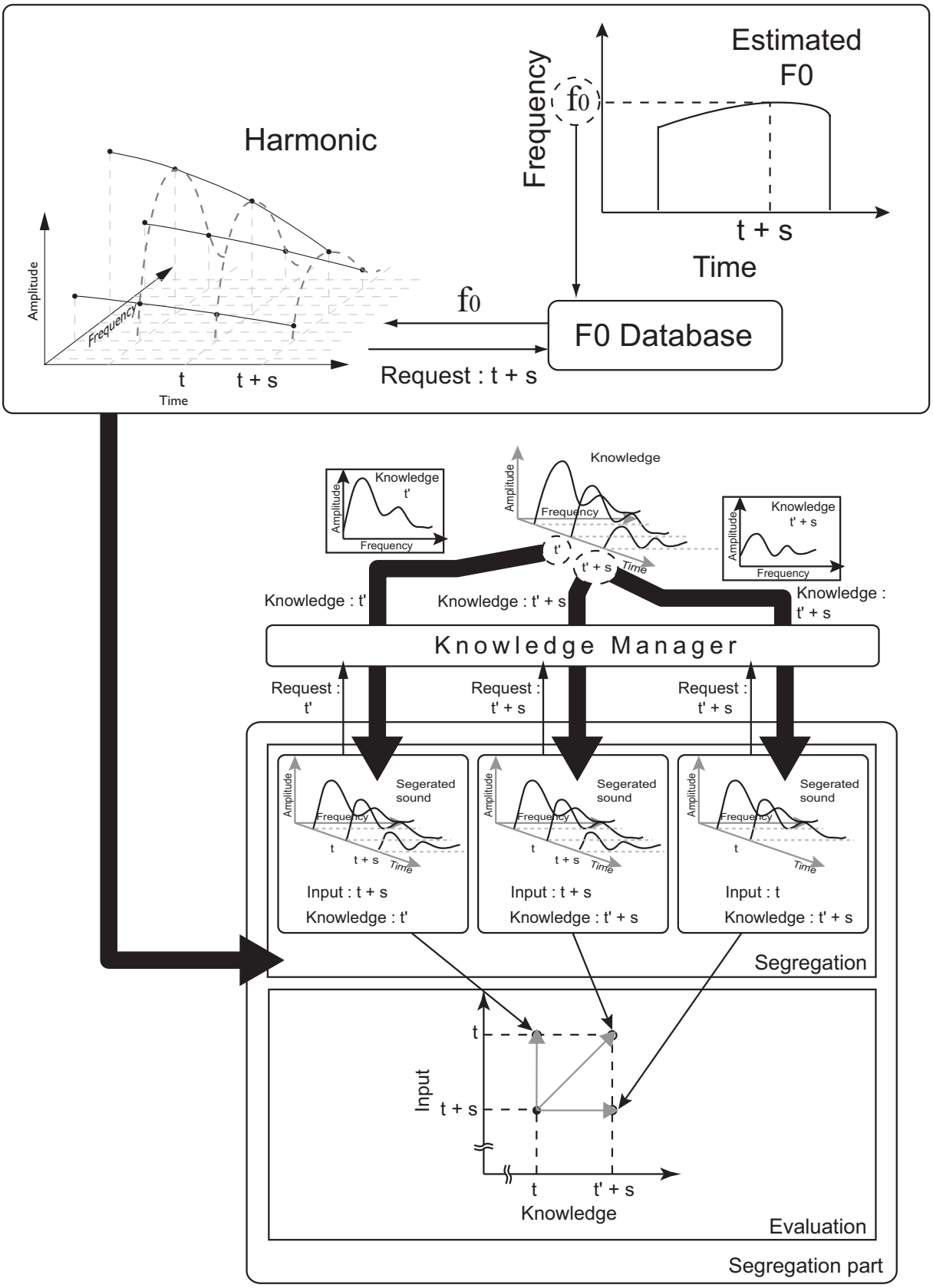


図 4.2 評価用モデルの処理の流れ

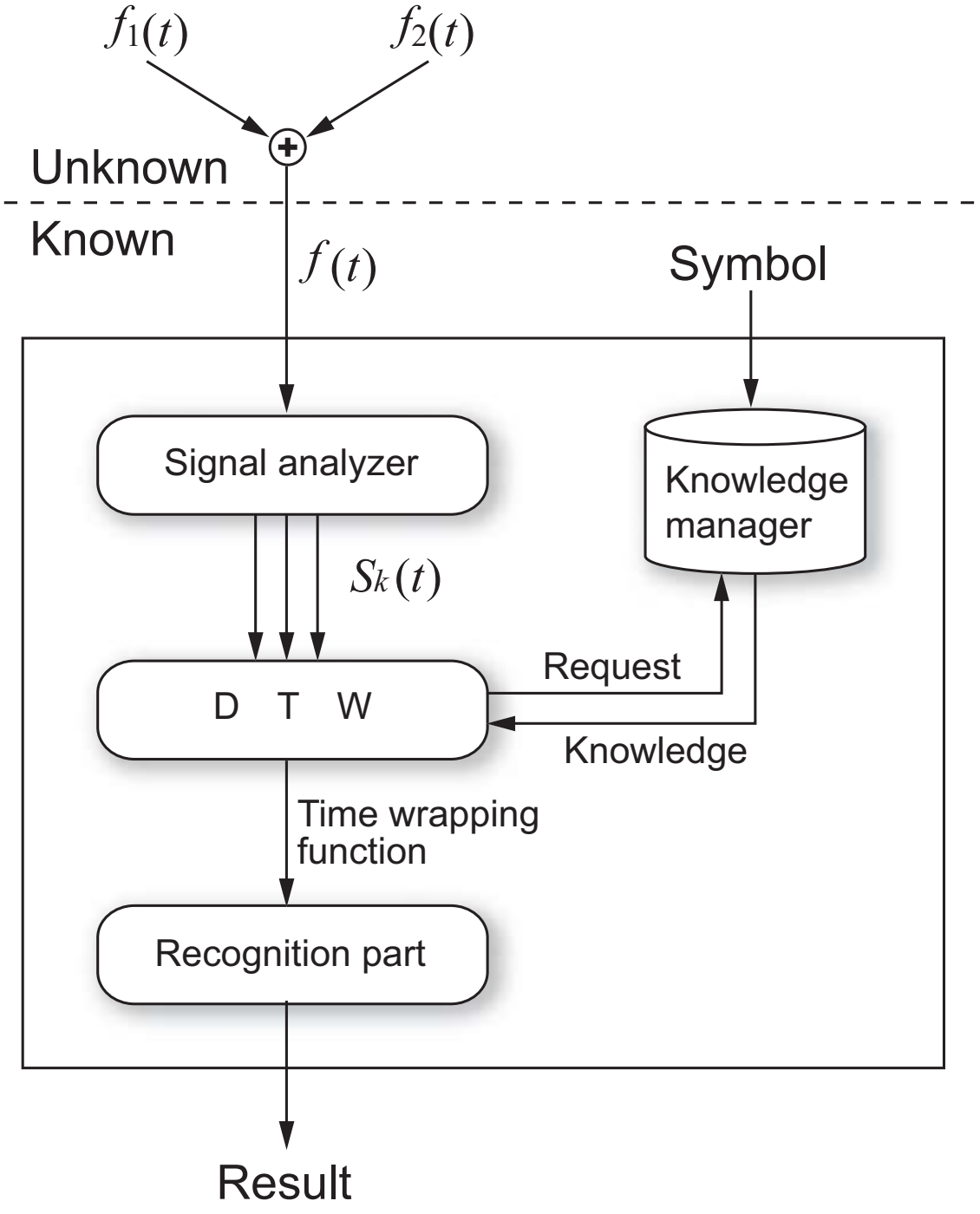
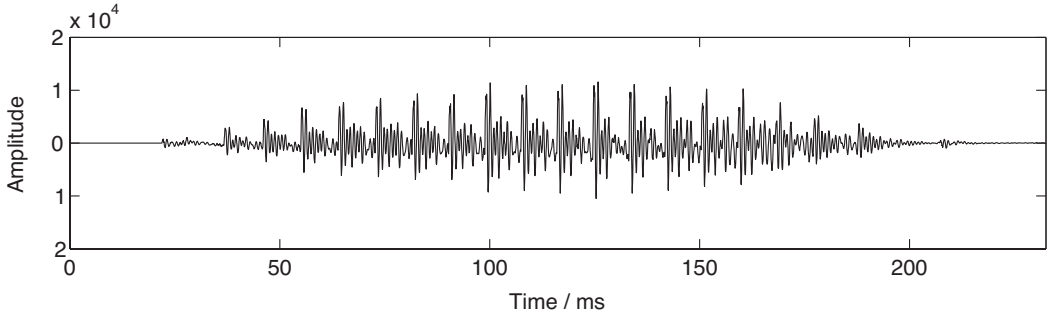
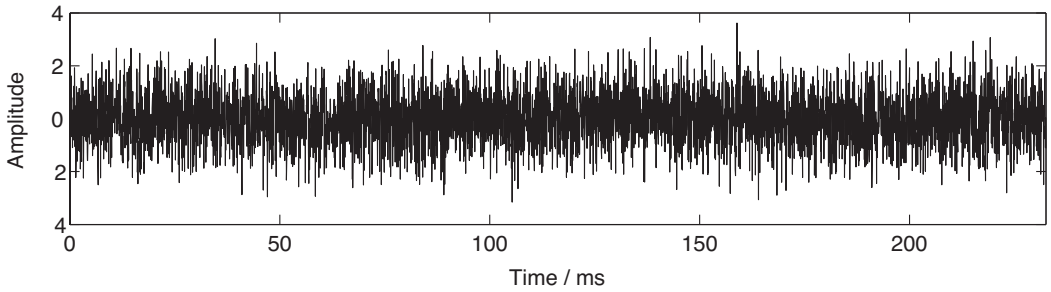


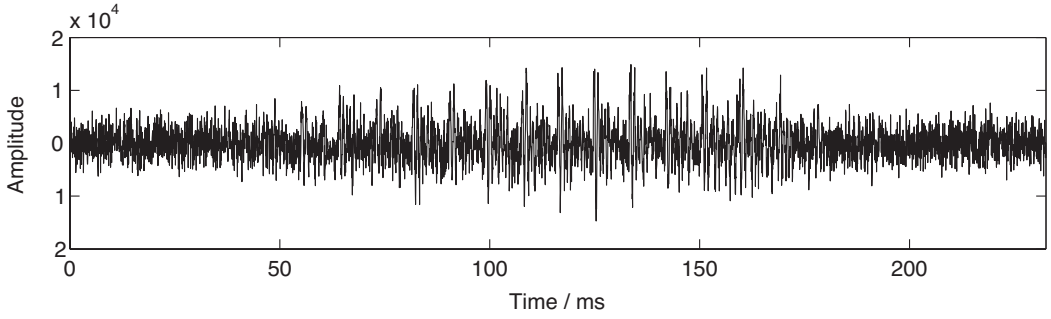
図 4.3 DTW による認識モデル



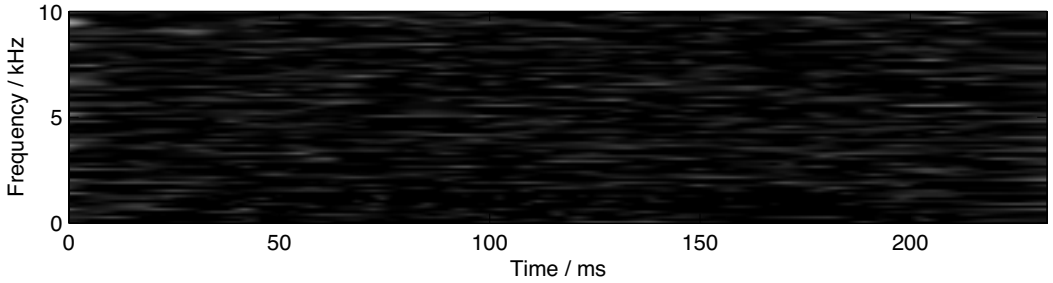
(a) クリーンな音声の時間波形 (mau, /a/)



(b) ガウス性白色雑音の時間波形



(c) 混合音 (SNR = 0 dB) の時間波形



(d) 混合音の瞬時振幅包絡 $S_k(t)$

図 4.4 入力音

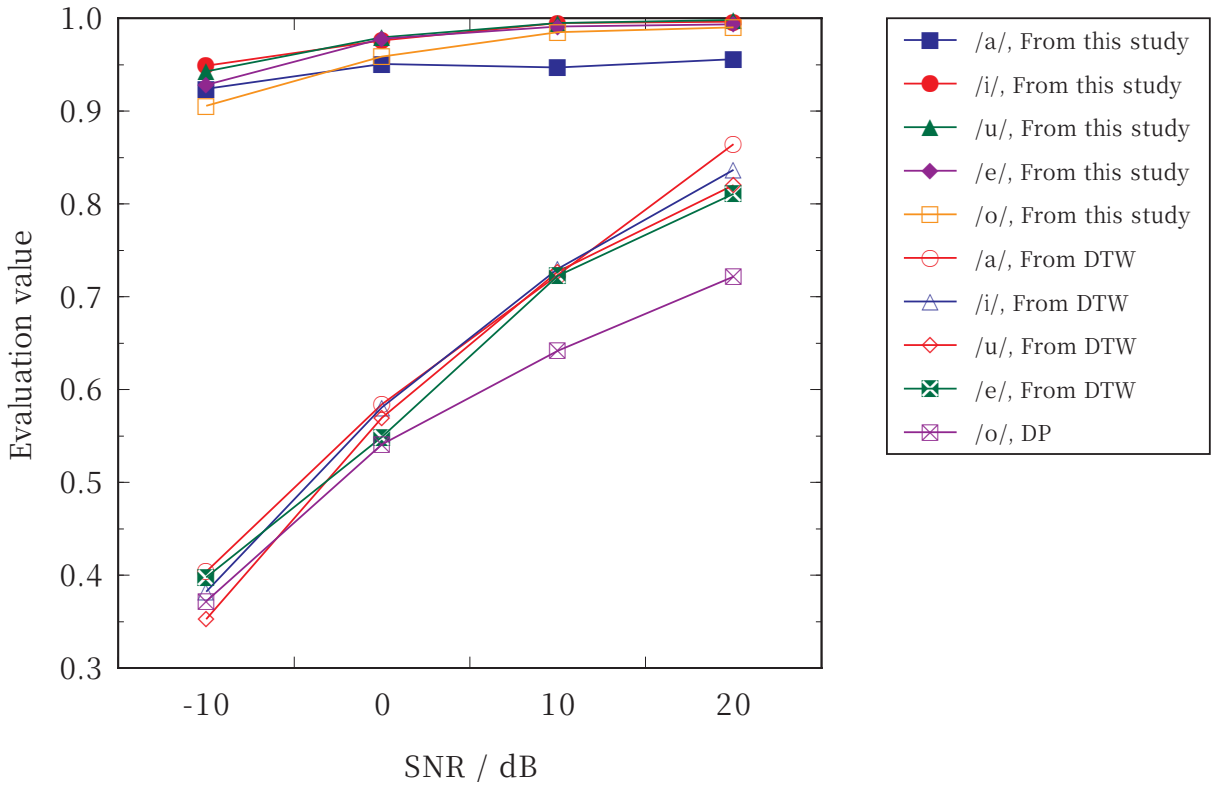


図 4.5 提案手法の評価におけるシミュレーション結果

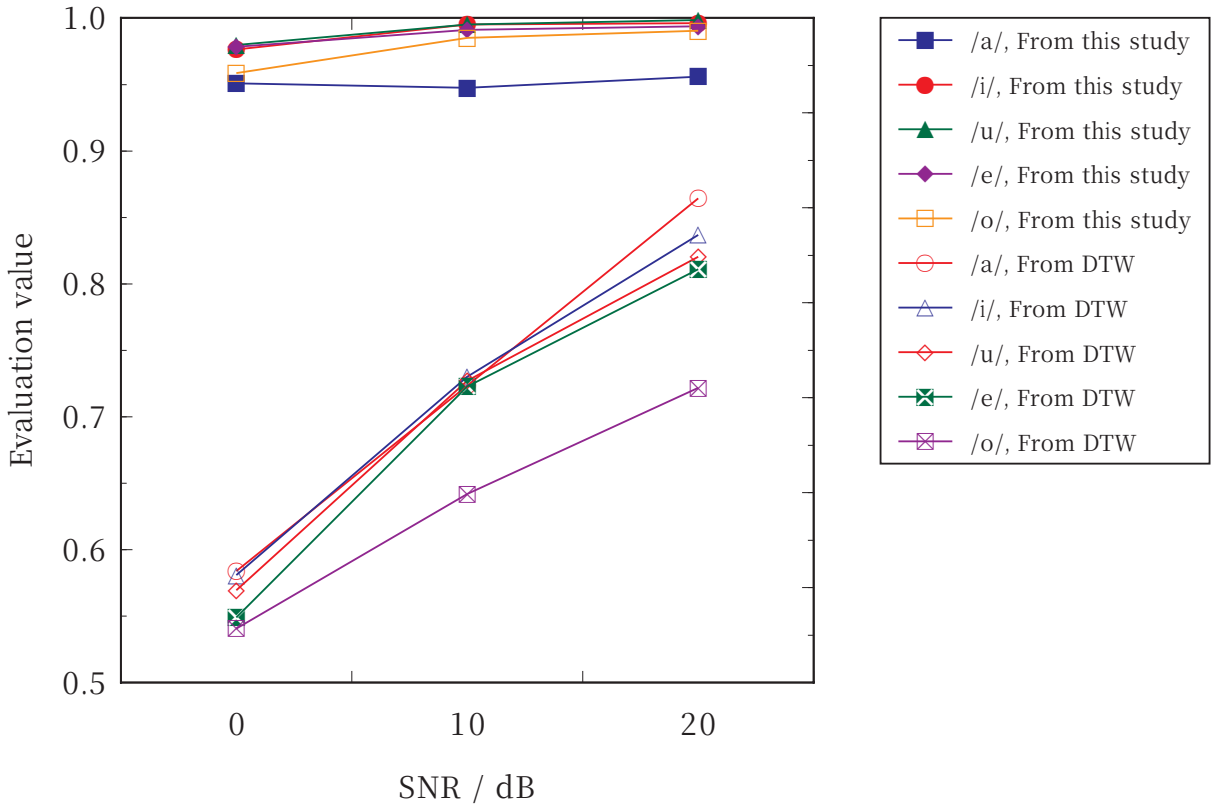


図 4.6 構築システムの評価におけるシミュレーション結果

第 5 章　結論

本章では、本論文の内容を要約し、今後の課題について述べる。

5.1　まとめ

本論文では、分離過程で目的音として妥当な音を分離しているかどうか、最終的な分離音が目的音として妥当であるかどうかを判断することにより、入力音中の目的音を認識する音源分離を認識の規範とした音声認識の手法を提案した。

はじめに、雑音環境下で音声を認識するという事は、i) 目的音と背景音が時間的にも周波数的にも予測不可能な重なりを持つ混合音から目的音を分離する、ii) 分離音の内容を認識するという 2 つの問題を解決する必要がある事を示した。また、認識の規範となりうる音源分離手法は、まずは分離の過程を議論する事ができる聴覚情景解析に基づく音源分離手法であるとし、さらにこの手法に知識を用いて積極的に目的音を分離するような手法を加えたものが認識の規範となりうる音源分離手法であるとした。そこで、音源分離処理と認識処理が融合した、音源分離を認識規範とする音声認識モデルを構築した。

提案した手法の有効性を示すために、白色雑音と単母音音声が混合した音から目的音声を認識するシミュレーションを行った。これにより、定常な白色雑音と単母音音声が空間的に加算された状況において本手法が有効である事を確認した。

5.2　今後の課題

今後、本研究の提案手法が単語音声に対応する事により、ワードスポッティングを用いた雑踏や自動車中での機械操作への応用などが考えられる。また、連続音声に対応する事により字幕放送でのリアルタイム字幕処理システム、さらに、不特定話者、自然発話音声に対応する事により機械と人間との間で音声による意思の伝達など多くの可能性を持っている。そこで、今後、本研究の提案手法が単語音声、連続音声に対応するために、必要な課題を以下

に示す。

様々な音声への対応

本研究では、連続音声の認識を最終的に目指している。そのためには、次に/aoi/や/ooi/のような母音のみからなる単語音声について、提案手法の有効性を確認する必要性がある。さらに、子音の認識を行う事が必要である。特に、破裂音や摩擦音のような無声子音は、F0 を持っていないために調波構造を利用した音源分離を行う事ができない。この点については新たな手法を提案モデルに対して加える必要がある。提案モデルでは知識を用いて積極的に目的音を分離する手法を用いている事から知識を有効に活用する事により無声子音の分離が可能になり、これにより子音の認識を行えると考えられる。

知識の改良

本研究の提案手法は音源分離を認識の規範としている。このため、音源分離時に欠かす事のできない知識は、提案手法では重要な位置を占めている。知識を積極的に用いる事により、提案手法は、単語音声、自然発話音声の認識時に問題となる調音結合や無声化などに対応しやすくなると考えられる。また、不特定話者への対応なども知識の活用により可能なのではないかと考えられる。このため、知識として数多くの音声をデータベースの形で保有する事が最初に考えられるが、現在のような単母音ではなく自然発話音声に対応するためには膨大な量の知識が必要であり、さらに不特定話者を考慮すると今回のようなデータベースタイプの知識には限界が見えてくる。よって、このようなデータベースではなく、音素表記を入力として受け取り、記号列の知識から、スペクトル包絡のような物理量からなる知識をその都度生成する仕組みが有効であると考えられる。これには、HMM (Hidden Markov Model) を用いて必要とする特徴量を生成する方法が考えられる。

システムの改良

今後、より複雑な状況に対応するためには、分離過程の妥当性を判断する部分が非常に重要になると考えられる。現時点では分離の妥当性を判断するために、DTW を用いてその妥当性を判断しているが 3.2.2 節で示しているように、これ以外の手法を用いることもできる。今後 DTW では対応できない状況となった場合には、状態遷移に確率を取り入れ HMM を用いることが考えられる。

さらに、前章の最後で示したように、本手法を有効に利用するためには現在用いている自

己相関法による F0 推定にかわって、雑音に対してよりロバストな手法を取り入れる必要がある。雑音に対してロバストな F0 推定法としては多くの研究がなされており [阿部 96, 阿部 00]、石本らが提案している PHIA [Ish01] などがあげられる。

短い時間で状態が変化する音声を認識するためには、セグメント処理が必要であり、リアルタイム処理を見据えた場合、それは絶対不可欠なものとなる。現時点では、提案モデルの多くの部分がセグメント処理化されているが、信号解析部などはバッチ処理を行うなど、まだ完全にセグメント処理化されている訳ではない。この部分は、最終的には改善しなければならない部分である。

ここで示した課題を克服することにより、本研究の提案手法は音声認識が利用できる状況を現在より大きく広げる可能性を持っている。

謝辞

本研究を遂行するにあたり、数多くの御指導と御鞭撻を賜りました赤木正人教授、党建武助教授に深く感謝の意を表します。日頃から熱心に御討論頂き、時には夜遅くまで御指導を賜り、有益な御助言を賜りました鵜木祐史助手に心から感謝いたします。本研究を客観的立場に立って数多くの有益な御指摘をいただき、また、公私にわたり数多くの御指導を頂いた伊藤一仁さんに厚く御礼を申し上げます。そして、日頃から数多くの議論と激励を頂いた赤木研究室の諸先輩方に厚く御礼を申し上げます、また、本研究の遂行にあたり多面にわたり御協力いただいた音情報処理学講座の皆様に感謝致します。

そして、最後に紆余曲折を経て今に至る私を温かく見守り、時には叱咤激励してくれた両親、友人に心からお礼を申し上げます。

参考文献

[Bre90] Bregman, A. S.: *Auditory Scene Analysis : The Parceptural Organization of Sound*, MIT Press, Cambridge, Mass., 1990.

[Bre93] Bregman, A. S.: Auditory scene analysis : hearing in complex environments, in McAdams, S. and E. Bigand eds., *Thinking in Sound: The Cognitive Psychology of Human Audition*, chapter 2, pp. 10–36, Oxford University Press, 1993.

[Bre94] Bregman, A. S.: 聴覚の情景分析とは, 音響学会誌, Vol. 50, No. 10, pp. 1007–1010, 1994, 河原　英紀　訳.

[Bro94] Brown, G. J. and M. Cooke: Comutational auditory scene analysis, *Comuter Speech and Language*, Vol. 8, No. 4, pp. 297–336, 1994.

[Coo01] Cooke, M. and D. P. W. Ellis: The auditory organization of speech and other sources in listers and computational models, *Speech Communication*, Vol. 35, No. 3-4, 2001.

[Dav52] Davis, K. H., R. Biddulph, and S. Balashek: Automatic recognition of spoken digits, *J. Acoust. Soc. Am.*, Vol. 24, No. 6, pp. 637–642, 1952.

[Ell94] Ellis, D. P. W.: A computer Implementation of Psychoacoustic Grouping Rules, *Proc. 12th Int. Conf. on Pattern Recognition*, 1994.

[Flo94] Flores, J. A. N. and S. J. Young: Continuous speech recognition in noise using sectral subtraction and HMM adaptation, *ICASSP*, Vol. I, , 1994.

[Ish01] Ishimoto, Y., M. Unoki, and M. Akagi: A Fundamantal Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency, *Poc. Eurospeech2001*, Vol. 4, pp. 2439–2442, 2001.

[Lip97] Lippman, R. P.: Speech recognition by machines and humans, *Speech Communication*, Vol. 22, pp. 1–15, 1997.

[McA93] McAdams, S. and E. Bigand eds.: *Thinking in Sound: The Cognitive Psychology of Human Audition*, Oxford University Press, 1993.

[Min92] Minami, Y., F. Martin, K. Shikano, and Y. Okabe: Recognition of noisy speech by

composition of hidden markov models, 信学技報 SP92-96, 1992.

[Shi01] Shimojo, S., C. Scheier, R. Nijhawan, L. Shams, Y. Kamitani, and K. Watanabe: 知覚モダリティを超えて：視知覚も及ぼす聴覚の効果, 音響学会誌, Vol. 57, No. 3, pp. 219–225, 2001.

[Tak01] Takiguchi, T., S. Nakamura, and K. Shikano: HMM-Separation-Based Speech Recognition for a Distant Moving Speaker, *IEEE Trans. on Speech and Audio Process.*, Vol. 9, No. 2, pp. 127–140, 2001.

[Vas97] Vaseghi, S. and B. Milner: Noise compensation methods for hidden Markov model speech recognition in adverse environments, *IEEE Trans. on Speech and Audio Process.*, Vol. 5, No. 1, pp. 11–21, 1997.

[阿部 96] 阿部, 小林, 今井：瞬時周波数に基づく雑音環境下でのピッチ推定, 信学論 (D-II), Vol. J79-D-II, No. 11, pp. 1771–1781, 1996.

[阿部 00] 阿部, 入野, 河原, 陸, 中村, 鹿野：調波成分の瞬時周波数を用いた基本周波数推定方法, 信学論 (D-II), Vol. J83-D-II, No. 11, pp. 2007–2086, 2000.

[鵜木 99] 鵜木, 赤木：聴覚の情景解析に基づいた雑音下の調波複合音の一抽出法, 信学論 (A), Vol. J82-A, No. 10, pp. 1497–1507, 1999.

[金田 97] 金田：騒音下音声認識のためのマイクロホンアレー技術, 音響学会誌, Vol. 53, No. 11, pp. 872–876, 1997.

[窪 02] 窪, 鵜木, 赤木：楽器音の音響的特徴を知識として利用した選択的分離抽出法, 聴覚研究会資料, Vol. 32, No. 10, 2002.

[古井 85] 古井 貞：デジタル音声処理, 東海大学出版会, 1985.

[小田 02] 小田：両耳処理の進化生物学, 音響学会誌, Vol. 58, No. 3, pp. 193–198, 2002.

[滝口 96] 滝口, 中村, 鹿野：雑音と残響のある環境下での HMM 合成によるハンズフリー音声認識法, 信学論 (D-II), Vol. J79-D-II, No. 12, pp. 2047–2053, 1996.

[中川 00] 中川：音声認識研究の動向, 信学論 (D-II), Vol. J83-D-II, No. 2, pp. 433–457, 2000.

[渡辺 96] 渡辺：音声認識システムの構築における諸問題とその解決, 信学論 (D-II), Vol. J79-D-II, No. 12, pp. 20022–2031, 1996.

[柏野 93] 柏野, 田中：モノラル楽器音の音源分離のための知覚的手がかりの検討と処理モデルの実装, 聴覚研究会資料, No. H-93-84, 1993.

[武田 97] 武田, 板倉：音源分離による音声認識性能の改善, 音響学会誌, Vol. 53, No. 11, pp. 883–888, 1997.

[郵政 99] 郵政省（編）：平成 11 年度版 通信白書, 1999.