論文

音韻知識にバイアスされた音韻推定：PHONOBEST ～現状と残された課題～

PHONOBEST: Phonetic Knowledge biased Estimation ～Current Status and Future Works～

勝瀬　郁代
Ikuyo MASUDA-KATSUSE

A new automatic speech recognition (ASR) method, PHONOBEST, is proposed, aiming to achieve robust ASR in a real environment. In ASRs, we need to compare observed speech with template speech in the pattern matching process. Such template speech can be regarded as a kind of top-down information generated from speech schema. In human speech perception, the top-down process seems to work more than preparing candidates for pattern matching. It also seems to predispose the segregation process to extract an object agreeing with the "expectation" generated by speech schema. PHONOBEST has introduced this active bias into the process of target speech estimation. A missing-data model is also implemented in PHONOBEST by using fine spectral representation with harmonic structures as template patterns; on the other hand, the spectral envelopes are supplied as template patterns in the prevailing ASRs. In this paper, the current status and remaining assignments are discussed.

キーワード：自動音声認識，音韻知識，トップダウン処理，音声知覚モデル，調波構造
Keywords: automatic speech recognition, phonetic knowledge, top-down process, speech perception model, harmonic structure

1．はじめに

音声は、人間にとって負荷の小さいコミュニケーションの手段であると考えられてきた。人間の発した音声を文字情報へ変換する(Speech To Text)ことができれば、人間の意図(希望する操作内容)を音声によってコンピュータに知らせることができる。そうすれば、人間が機械を操作する際の負担を軽減することができると信じられ、音声認識技術の研究開発は進められてきた。そして近年、自動音声認識システムは、大語彙連続発話音声の認識率が95%を超えるに至っており、パーソナルコンピュータで実時間動作が可能となっている。現在の音声認識技術は、人間の聴覚の性質を利用している部分はあるものの、基本的には、音響学などの物理的理論や、確率論などの数学的理論に基づくものであり、人間が音声を認識するメカニズムとは、本質的に異なる方式で認識を行っている。そして、この大量のデータと計算機の能力をバックに"力技"で認識を行う方法は、あるところまでは成功してきたといえる。

それにも関わらず、我々の身近な機械に、音声認識機能を有するものがそれほど多くない。表向きの数値(認識率)の割に実用化例が少ない原因はどこにあるのか。一般に報告されているパフォーマンスは、背景騒音のない環境において、ドメインの限られた話題の、さらに限られたタスクを、文書を読み上げるがごとく明瞭に発声した場合の数値である。残念ながら、自動音声認識技術の応用が望まれる場面では、このような条件は必ずしも満たされないのである。これまでの音声認識の研究開発スタイルは、まずは理想的な環境を前提に技術を構築し、実環境に起因するさまざまな問題に対しては、それから対応していくものであった。このスタイルでは、場合によっては深刻なスケールアップ問題に直面することがあるが、力技でいけるところまでいって、解決できず残された部分に対して、人間のメカニズムの解明に基づく方法などの必要性が高まってくるものと考えられてきた[3]。

このような背景から、本研究では、主に騒音環境下で有効に働くと考えられる、人間の音声知覚のための戦略を取り入れつつ、現在主流となっている確率統計的アプローチと整合する、新しい音声認識手法を提案し、いくつかの課題について言及する。

近畿大学産業理工学部　情報学科　産業技術研究科　経営工学専攻　katsuse@fuk.kindai.ac.jp

2. 提案手法の特徴

本手法は、これまでの音声認識システムと比較して2つの異なる特徴を持つ。一つは、学習音声データによって構築された音韻知識を入力信号と照合するためのデータとして用いるのではなく、観測信号におけるある特定のスペクトルパターンの存在の「期待」として利用することである。これは、人間の音声知覚過程における能動的トップダウン処理の導入に相当する。もう一つの特徴は、観測信号のスペクトル表現方法である。従来音声認識システムでは音声情報はスペクトルの概形で表現されている。本提案手法では音響分析段階で周波数分解能を高くとることで、より詳細なスペクトル表現を用いている。

2.1　概念駆動型処理の実現

人間の音声言語理解の過程を計算機による情報処理のアナロジーで捉えると、音声信号の感覚的変換物をその信号に意味を与える概念過程に連結するための一連の認知過程であるといえる。そしてこのような一連の認知過程はボトムアップ処理(データ駆動型処理)とトップダウン処理(概念駆動型処理)に大別される。ボトムアップ処理は、外界の情報を直接知覚対象に変換する処理であり、トップダウン処理は、可能な解釈についての知識、すなわちスキーマがその事物の知覚を助ける時、そこで起こっている処理である。音声言語理解の過程では、概念化において音韻論的、統語論的、及び意味論的知識が積極的に利用されていると考えられている。

このように、人間の音声知覚過程では、トップダウン処理とボトムアップ処理はともに貢献するが、トップダウン処理は、ボトムアップ処理の結果の不完全な部分を補うという受動的な働きをするだけではない。むしろ、ある特定の音声の存在に関する「期待」を生成し、その期待に合った答えを得られるように処理機構にバイアスをかけるという能動的な働きをするものである。そして、トップダウンの働きは、騒音下での音声知覚に対して重要な役割を果たしている。

本研究では、音声知識に基づく能動的トップダウン処理を導入することによって、音声と騒音が混在する観測信号から音声の特徴量を推定し、騒音下の音声認識を実現するシステムを提案する。本稿では、音声知識として音韻知識を用いることから提案手法をPHONOBEST(PHOnetic kNOwledge Biased ESTimation)と呼ぶ。

2.2　知識との照合に適切な音声表現

一般に、音声周波数分析手段として、ケプストラム分析、LPC分析などが用いられる。自動音声認識システムでは一般に、このようなスペクトル分析手法を用いて観測された音声信号から調音フィルタの振幅伝達特性を抽出し、音声の標準的な音響特徴と比較照合することが行われている。このように、音声に関する事前知識との照合の過程における音声表現としてスペクトルの概形を用いることで、韻律の違いによって生じる微細なスペクトル構造の違いを吸収することには成功した。しかし聴覚による音声知覚モデルとして妥当かどうかは別問題である。聴覚末梢における情報表現から生成可能なスペクトル概形からでは母音の弁別が行えない場合があることがよい例である。話し手から聞き手へ伝えられる音声情報表現は、声道伝達特性自体ではなく、時間周波数表現の曲面に表現される伝達特性を声帯音源による周期的駆動により標本化して抽出したものであると解釈できる。しかし、ホルマントなどの母音的特徴を保持するように声道伝達特性を再表現するには、声道音源の駆動周期は短すぎる(つまり基本周波数は高すぎる)のである。de Cheveigné と Kawaharaは、このような現象を説明することができる母音同定のmissing-dataモデルを提案している[6]。Fig. 1は、missing-dataモデルの概念図である。スペクトル包絡の知識を持つが、聴取された音響信号と照合する際には、基本周波数の倍音付近に限定した領域でのマッチングによるパターン認識処理となっている。このように、声道伝達特性が音声の音韻性を与える重要な特徴量であるにも関わらず、その声道伝達特性を時間方向と周波数方向の両方で、声帯振動周波数によって決まる間隔で標本化してから聞き手へ伝えるというのは一見無駄な操作を行っているように見える。しかし、騒音環境での音声情報伝達を考えるとこれは実に合理的な戦略であるように思える。ある規則性を伴って情報が局在しているということは、同時に存在し、かつその規則に従わない情報からポップアップしやすいからである。聴覚が騒音環境下でターゲットとなる音を選択的に聴取する能力について、Bregmanはその著書"Auditory Scene Analysis(聴覚情景分析)[4]"にまとめている。そして、基本的な原理の中でも特に頑健に働く重要な原理として「調波性の原理」をあげている。スペクトル表現を包絡のみに要約する従来の音声認識では、音声情報が持つこの貴重な手がかりを放棄しているともいえる。本研究では、知識を用いた音韻推定を行う際に、従来の音声認識システムのようにスペクトルの概形をそのまま用いるのではなく、de Cheveignéらのモデルのように、スペクトル概形のうち倍音付近に限定した領域のみを用いて音韻推定を行う。

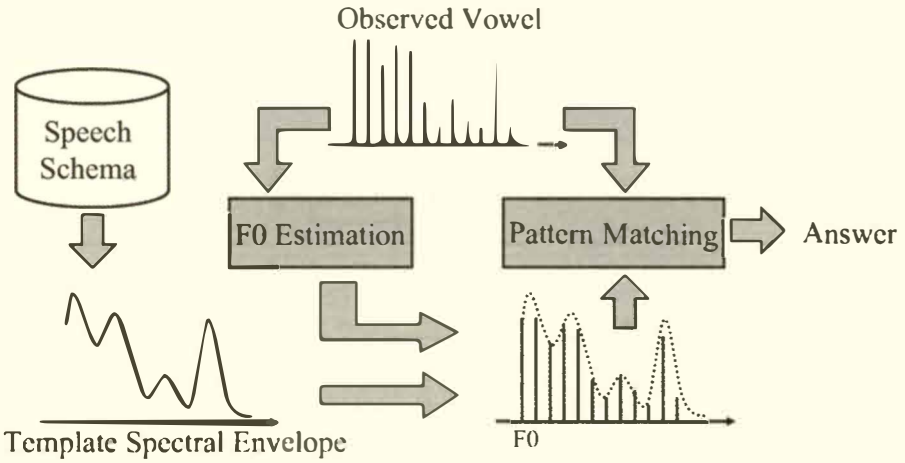

Fig. 1. Missing-data model

3. システムの概要

PHONOBESTは、音楽信号のメロディとベースラインを推定する手法であるPreFEst [9][10]の拡張として実現されている。しかしながら、本システムは、従来のHMMに基づく音声認識システムの一部として組み込まれる形で実現できるという特徴をもつ。本節では、離散HMMに基づく単語認識システムを適用した場合の実装方法について述べる。

3.1 離散HMMに基づく単語音声認識の枠組み

Fig. 2は離散HMMに基づく音声認識の概略を示している。はじめに、短時間周波数分析により入力音声信号のスペクトル包絡が得られる。得られたスペクトル包絡は、符号帳を用いてベクトル量子化される。これが入力信号と標準信号の照合段階に相当する。量子化された特徴ベクトル(セントロイド)列が単語HMMから出力されたとみなし、最大の確率を与えるHMMを持つ単語が認識結果として採用される。

3.2 新規部分の概要

Fig. 3は、PHONOBESTにおいて、観測信号からセントロイド列が求められるまでの流れを示したものである。Fig. 2の破線で囲まれた部分をFig. 3に置き換えることにより、本提案手法を離散HMM型音声認識システムに組み込むことができる。本節では、Fig. 3で示された提案部分の概略を述べる。

観測信号を周波数分析し、得られた周波数スペクトルを確率密度関数 $p_{obs}^{(t)}(x)$ として見なす。そして $p_{obs}^{(t)}(x)$ が音モデル$p(x|F,m,c(h|F,m))$の重み付き混合分布から生成されたと見なす。混合重み値を$w^{(t)}(F,m)$とする。そして、最大事後確率推定により、$c_0^{(t)}(h|F,m)$, $w_0^{(t)}$が音モデルのパラメータ$c^{(t)}(h|F,m)$と重み値 $w^{(t)}(F,m)$ の最も起こりやすいパラメータとして与えられた時の $w^{(t)}(F,m)$ を推定する。パラメータ$c_0^{(t)}(h|F,m)$ は符号帳のセントロイドの振幅包絡から生成される。推定された$w^{(t)}(F,m)$ の値から、符号帳のセントロイドの「音韻ドミナンス」を各時刻で定義することができ、最大の音韻ドミナンスを持つセントロイドを各時刻で出力することにより、セントロイド列を得ることができる。以下の各節では、各処理部に分けて説明する。

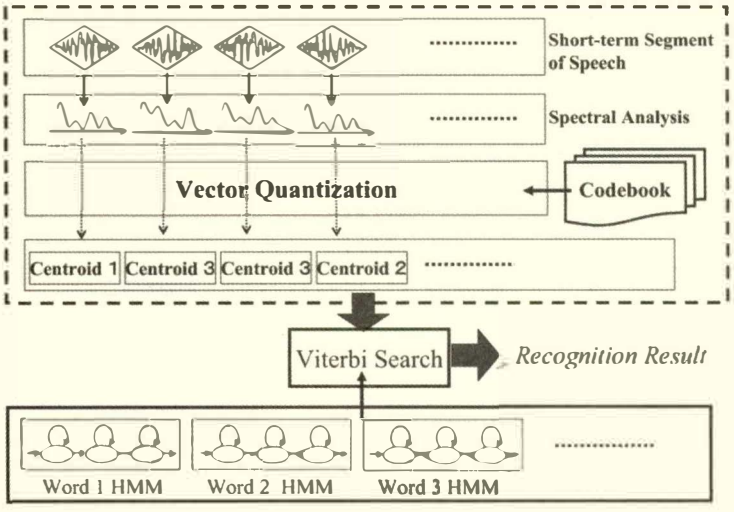

Fig. 2. Scheme of discrete HMM based recognition.

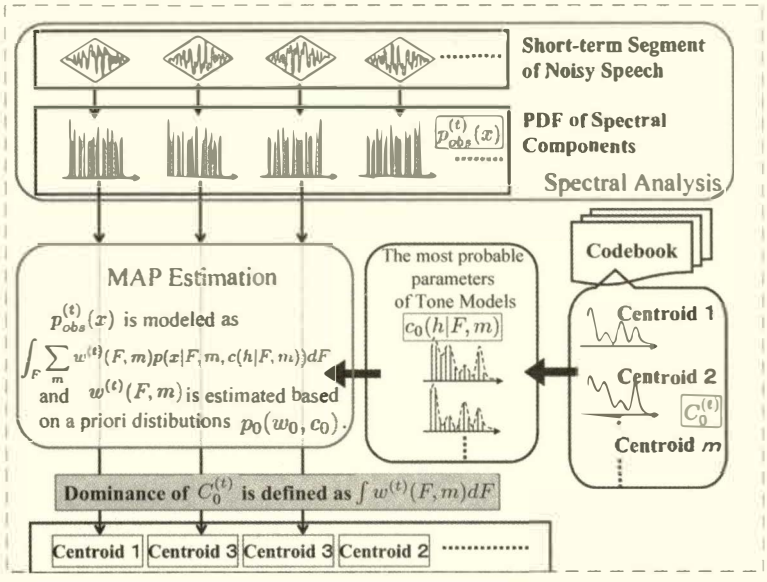

Fig. 3. Flow of the process to obtain a sequence of centroids in PHONOBEST.

3.3 周波数分析過程

観測信号を高域強調し、次式で定義される時間窓によって切り出された短時間信号をフーリエ変換する。時間窓関数は、がボール関数に二次のカーディナルスプライン関数を畳み込んだ次の関数とする。

$$w_p(t) = exp\left(-\pi\left(\frac{t}{t_0}\right)^2\right) \odot h\left(\frac{t}{t_0}\right) \tag{1}$$

$$h(t) = \begin{cases} 1-|t| & |t|<1 \\ 0 & otherwise \end{cases} \tag{2}$$

t_0は信号の基本周期である。時間窓は、観測信号の基本周期の関数となっているが、観測信号の基本周期は予めわかっていないのが普通である。幸い、窓長が実際の周期の数割程度のずれなら問題なく瞬時周波数が求められることがわかっている[2]。基本周波数のおおよその値は阿竹ら[2]の手法等により求めることができるが、本研究では、目的音声の基本周波数の平均をおおよその値として予め与えている。

次に、フィルタバンクの出力から瞬時周波数を求める。フィルタバンクの出力を $X(\omega,t)=a+jb$ とすると、その瞬時周波数は、

$$\lambda(\omega,t)=\omega+\frac{a\frac{\partial b}{\partial t}-b\frac{\partial a}{\partial t}}{a^2+b^2} \tag{3}$$

で与えられる。求めたい周波数成分の候補はフィルタの中心周波数から瞬時周波数への写像の平衡点となり、次のように求めることができる[15]。

$$\psi(t)=\omega|\lambda(\omega,t)-\omega=0,\frac{\partial}{\partial\omega}(\lambda(\omega,t)-\omega)<0 \tag{4}$$

周波数成分のパワーは、短時間フーリエ変換のパワースペクトルの値として得られるので、

$$\psi(\omega,t)=|X(\omega,t)| \tag{5}$$

となる。

3.4　推定過程

PreFEstでは、調波構造をもったスペクトル成分の集合である音モデルを仮定している[9]。基本周波数がFであるm番目の音モデルの確率密度関数を以下のように表す。

$$p(x|F,m,\mu^{(t)}(F,m))=\sum_{h=1}^{H}p(x,h|F,m,\mu^{(t)}(F,m)) \tag{6}$$

ここで、

$$p(x,h|F,m,\mu^{(t)}(F,m))=c^{(t)}(h|F,m)G(x|F,h) \tag{7}$$

であり、 Hは考慮する倍音の数、G は$F\cdot h$ で最大値を持つガウス分布である。

前節で求めた観測信号のスペクトルを確率密度関数とみなし、この確率密度関数をすべての可能な基本周波数の音モデルの重み付き混合モデルから生成されたとみなす。

$$p(x|\theta^{(t)})=\int_{Fl}^{Fh}\sum_{m=1}^{M}w^{(t)}(F,m)p(x|F,m,\mu^{(t)}(F,m))dF \tag{8}$$

$$\mu_0^{(t)}(F,m)=\{c_0^{(t)}(h|F,m)|h=1,\cdots,H\} \tag{9}$$

ここで、F_lとF_hはそれぞれ基本周波数の下限と上限を示す。重み値$w^{(t)}(F,m)$ はそれぞれの音モデルの重み値である。我々は今、事前の知識として $w^{(t)}(F,m)$ と $c^{(t)}(h|F,m)$ の最も期待されるパラメータ$w_0^{(t)}(F,m)$と$c_0^{(t)}(h|F,m)$ を持つ。そしてこのパラメータから生成された事前分布に基づいてモデルのパラメータ $w^{(t)}(F,m)$ と $c^{(t)}(h|F,m)$ を推定したい。事前分布に基づくパラメータの最大事後確率を推定するには、EMアルゴリズムの繰り返し演算において古いパラメータを新しいパラメータに更新すればよい[10]。この時、新しいパラメータ とは無情報事前分布が与えられた時の最尤推定値 $w_{ML}^{(t)}(F,m)$ と$c_{ML}^{(t)}(h|F,m)$を用いて以下のように求められる。

$$\overline{w^{(t)}(F,m)}=\frac{\overline{w_{ML}^{(t)}(F,m)}+\beta_\omega^{(t)}\omega_0^{(t)}(F,m)}{1+\beta_\omega^{(t)}} \tag{10}$$

かつ、

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{\omega_{ML}^{(t)}(F,m)c_{ML}^{(t)}(h|F,m)} + \beta_{\mu}^{(t)}(F,m)c_0^{(t)}(h|F,m)}{\overline{\omega_{ML}^{(t)}(F,m)} + \beta_{\mu}^{(t)}(F,m)} \tag{11}$$

であり、ここで、

$$\overline{w_{ML}^{(t)}(F,m)} = \int_{\infty}^{\infty} p_{\psi}^{(t)}(x) \frac{w'^{(t)}(F,m)p(x|F,m,\mu'^{(t)}(F,m))}{\int_{Fl}^{Fh}\sum_{v=1}^{M} w'^{(t)}(u,v)p(x|u,v,\mu'^{(t)}(F,v))du} dx \tag{12}$$

かつ、

$$\overline{c_{ML}^{(t)}(F,m)} = \frac{1}{\overline{w_{ML}^{(t)}(F,m)}} \int_{-\infty}^{\infty} p_{\psi}^{(t)}(x) \frac{w'^{(t)}(F,m)p(x,h|F,m,\mu'^{(t)}(F,m))}{\int_{Fl}^{Fh}\sum_{v=1}^{M} w'^{(t)}(u,v)p(x|u,v,\mu'^{(t)}(F,v))du} dx \tag{13}$$

である。このように、$w^{(t)}(F,m)$は$w_{ML}^{(t)}(F,m)$と$w_0^{(t)}(F,m)$の重み付き平均として、$c^{(t)}(h|F,m)$は$c_{ML}^{(t)}(h|F,m)$と$c_0^{(t)}(h|F,m)$の重み付き平均として更新される。$w_0^{(t)}(F,m)$は音モデルの重み値であるので、その値は、基本周波数に関しては韻律の知識によって、スペクトルに関しては言語的な知識によって制御可能である。同様に、$c_0^{(t)}(h|F,m)$は音モデルの成分の相対的振幅を表しているので、その値は音韻的知識によって制御可能である。例えば、パラメータ$c_0^{(t)}(h|F,m)$は、事前知識として持っている標準音声のm番目の標準音声のスペクトル包絡$C_0^{(t)}$から次のように生成される。

$$c_0^{(t)}(h|F,m) = \frac{C_0^{(t)}(x|m)\delta(h\cdot F)}{\sum_{h=1}^{H} C_0^{(t)}(x|m)\delta(h\cdot F)} \tag{14}$$

さらに、β_wとβ_μは期待の強さを表していると解釈できる。

3.5　セントロイドの選択

求めたいのは音モデルそのものではなく、音モデルの生成のために用いられた標準音声のスペクトルの存在の妥当性である。m番目の標準音声のスペクトルの時刻tにおける音韻ドミナンス$PHD^{(t)}(m)$を以下のように定義する。

$$PHD^{(t)}(m) = \int w^{(t)}(F,m)dF \tag{15}$$

時刻tで最大の音韻ドミナンスを持つセントロイドをその時刻の出力とする。

4. 評価

PHONOBESTの現時点での認識性能を確認するために、以下の条件で単語認識実験を行った。使用した音声認識エンジンは、HTK ver. 3.3[11]である。ベクトル量子化のコードブックサイズは32であり、HMMの状態数は13(ただし両端を除く)である。PHONOBESTの基本周波数の範囲は90～170Hzとし、4Hzステップとした。標本化周波数は8kHzであり、FFTは1024ポイントである。学習ならびにテスト音声は産総研単語音声データベース[17]の話者S0001, S0003, S0010の最初の100単語を用いた。認識率は、話者S0001は81%、S0003は79%、S0010は86%であった。なお、従来手法(HTK単独)では、100%の認識率が出ている。

5. PHONOBESTの利点

PHONOBESTは、現在のところ、従来手法よりもそのパフォーマンスは低いが、次章に挙げるいくつかの課題を克服できれば、特に騒音環境下においてその有効性が期待される。これまで騒音環境での頑健な音声認識システムの実現を目指して多くの手法が取り組まれてきた。従来の研究は、入力信号(音声＋ノイズ)と照合信号(学習音声により構築された音響モデル)のミスマッチを避けるための工夫を行っている。それらは大きく分けて次の2つのアプローチに分けることがで

きる。一つは、入力信号から目的音声のみの特徴量を抽出することにより、入力信号の特徴ベクトルと照合信号の特徴ベクトルをマッチさせるという戦略をとるものである。CMN (Cepstrum Mean Normalization) [1]、スペクトルサブトラクション[16]、櫛形フィルタ[8]、独立成分分析などはこのアプローチに属する。ノイズの性質によっては抽出された音声信号またはその特徴量は音響的歪みを受けることがあり、照合信号とのミスマッチの原因となっている。もう一つは、音響モデルのノイズ補償を行うアプローチである。このアプローチでは照合段階で利用される音響特徴量は目的音声信号と騒音が重畳されたものである。つまり、照合パターンを音声と騒音が重畳させて生成することにより、入力信号の特徴ベクトルと照合信号の特徴ベクトルをマッチさせるという戦略をとるものである。PMC(Parallel Model Combination) [7]などがこのアプローチに属し、クローズドな評価実験において比較的高い性能を示している手法である。しかし、より一般的な応用を考えれば、背景騒音の性質やS／N比に仮定を一切設けないことがより頑健性を保証することにつながるものと考える。本手法は、そもそも観測信号のそのままを照合対象としないため、背景騒音に関する仮定は必ずしも必要でないという利点がある。

6. 残された課題と解決策

6.1　無声子音の取り扱い

本システムでは音声の調波構造を仮定したモデルを用いている。そして、本来は調波構造を持たない無声子音部でも同じモデルを適用している。このことが、背景騒音がない場合に本システムのパフォーマンスが従来方法より低い理由の一つになっている。音声包絡を音響情報として用いている一般的な音声認識の枠組みでは、調波構造を持つ有声部と、調波構造を持たない無声部は、特に別々にモデリングする必要はないが、本システムでは、異なる取り扱いが必要である。しかしながら、無声子音のモデリングを別に行う必要があるかどうかは疑問である。背景騒音がある場合、無声子音部は、有声部に比べて相対的にS／N比が小さくなる。そのため、仮に無声子音のモデリングを行ったとしても、実際上はあまり効果が得られない可能性がある。一方で、無声子音の前後の有声部では、「わたり」と呼ばれる、特徴的な音響的変化が見られることが知られている。人間の音声知覚においても、この「わたり」が子音同定の重要な手がかりになっていることが知られている。よって、無声部の子音の同定は、子音前後の有声部が与える情報で補うこととし、無声部の推定結果は認識誤りをもたらすものとして、認識結果に反映されにくいようにすることも考えられる。例えば、推定された$w(F,m)$の値をViterbiの尤度更新計算時に反映させる方法があろう[14]。

6.2　スペクトルの推定精度と照合

音響分析過程において、瞬時周波数から周波数成分の周波数と振幅を決定する際、フィルタの出力に主要な1つの成分があるときには安定した平衡点が得られるが、複数の倍音成分が混在していたり、ノイズ成分が多い場合には、安定した平衡点を得ることができない。ところで、フィルタバンクにノイズ成分が含まれるということは、複数成分による周波数や振幅の変調が観測されるはずである。この変調の大きさを調べれば、そのフィルタから得られたスペクトル成分の推定精度を知ることができる。Cohen[5]は、周波数の広がりを示す指標として帯域幅Bを考え、次のように定義した。

$$B^2 = \int \left(\frac{A\prime(t)}{A(t)}\right)^2 A^2(t)dt + \int (\phi\prime(t) - <\omega>)^2 A^2(t)dt \tag{16}$$

ここで、$A(t)$は振幅、ϕは位相である。また$<\omega>$は平均周波数で、次のように求められる。

$$<\omega^2> = \int \left(\frac{A\prime(t)}{A(t)}\right)^2 A^2(t)dt + \int \phi\prime^2(t)A^2(t)dt \tag{17}$$

上記式で求められた帯域幅の値が大きければ、そのフィルタバンクの出力は精度が低いことがわかる。また、得られた周波数成分、または周波数帯域それぞれに、精度を得ることができる。そして、この精度が低い場合、その成分や帯域を除いた周波数構成で確率密度関数を仮定し、照合を行えば、騒音の影響を受けにくい周波数帯域のみの利用が可能となる。このようなことは、スペクトル包絡でマッチングを行う従来方法では簡単にはできないことであり、PHONOBESTの大きな特徴となる。

6.3　事前分布の積極的な利用

騒音下では、スペクトル成分が振幅変調と周波数変調を受け、結果的に$w(F,m)$の推定値が影響を受ける。振幅変調による影響はセントロイドの選択誤りにつながり、周波数変調による影響は基本周波数の推定誤りにつながる。これらの誤りは同じレベルで起こるのではなく、セントロイドの選択誤りよりも、基本周波数の推定誤りの方が小さく現れる傾向が見て取れる。そのため、現在は、$w(F,m)$からセントロイドの推定と基本周波数の推定を同時に行っているが、まずは、基本周波数の推定を行い、推定された$w(F)$を、$w(F)$の最も起こりやすいパラメータ$w_0(F)$として再度事前分布を構成し、$w(F,m)$ の再推定を行うことで、推定精度は向上すると思われる。

そもそも $w_0(F,m)$ は種々の言語知識によって生成される「期待」を反映できるパラメータである。最も反映させやすい知識は韻律に関するものであろう。日本語単語知覚において、様々な音声情報のうち、言語情報を与えるものは音韻情報であると見なされがちである。これまで韻律情報は主にパラ言語情報を担うという観点から研究されることが多かった。実際に、韻律が不適切であっても単語同定能力にはほとんど違いがみられない。一方で、我々人間は様々な言語的知識によって予測されやすい言葉ほど騒音下での検出能力が高いことが示されている[12]。さらに勝瀬[12]は、日本語単語知覚における単語の韻律情報(ピッチアクセント型)の効果について調査している。そして、復唱課題においては、単語アクセント型の適切性と復唱を開始するまでの反応時間の間には、負の相関があることがあることを明らかにしている。また、有意味単語の知覚であっても、単語アクセント型が適切でない場合は、アクセント型が適切な有意味単語の知覚時における脳反応よりも、無意味単語知覚時の脳反応に近い可能性すら、示唆されている[13]。このように、非騒音下における我々人間の“Speech to Text”では、韻律情報がなくても“Text”は十分得ることができるにも関わらず、“Text”に至るまでのプロセスには韻律情報が関わっているという事実は、韻律が音韻同定のための重要な手がかりを与えるパラメータである可能性を示唆している。PHONOBESTには韻律情報に基づく事前分布を構成する枠組みが最初から用意されており、韻律が音韻同定に貢献することを工学的な立場から証明することができるかもしれない。

参考文献

[1]B. Atal, “Effectiveness of Linear Prediction Characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am., vol.55, pp.1304－1312, 1974.
[2]阿竹義徳, 入野俊夫, 河原英紀, 陸金林, 中村哲, 鹿野清宏, “調波成分の瞬時周波数を用いた基本周波数推定法,” 電子情報通信学会論文誌 D-II, Vol. J-83-D-II No. 11, pp. 2077－2086, 2000.
[3]安藤彰男著, リアルタイム音声認識　第1章 音声認識入門, 電子情報通信学会, 2005.
[4]A. Bregman, Auditory Scene Analysis, Mit Press, 1990.
[5]L. コーエン(吉川昭, 佐藤俊輔訳), 時間－周波数解析, 朝倉書店, 1998.
[6]A. de Cheveigné and H. Kawahara, “Missing-data model of vowel identification,” J. Acoust. Soc. Am., vol.105(6), pp.3497－3508, 1999.
[7]M.J.F. Gales and S.J. Young, “Robust Continuous Speech Recognition using Parallel Model Combination,” IEEE Transactions on Speech and Audio Processing vol.4, 1996.
[8]J. T. Graf and N. Hubing, “Dynamic time warping for the enhancement of speech degraded by white Gaussian noise,” Proc. Int. Conf. Acoust. Speech Signal Process.(ICASSP), II, pp.339－342, 1993.
[9]後藤真孝, “音楽音響信号を対象としたメロディーとベースの音高推定,” 電子情報通信学会論文誌 D-II, vol.J84(I), pp.12－22, 2001.
[10]M. Goto, “A predominant-F0 estimation for CD recordings: MAP estimation using EM algorithm for adaptive tone models”, Proc. of ICASSP2001, 2001.
[11]T. Hain, P.C. Woodland, G. Evermann, and D. Povey, “The CU-HTK March 2000 HUB5E Transcription system,” Proc. Speech Transcription Workshop 2000, 2000.
[12]Ikuyo Masuda-Katsuse, “Contribution of pitch-accent information to Japanese spoken-word recognition,” Acoust. Sci. & Tech., 27(2), 97－103, 2006.

[13]勝瀬郁代, 志方太一, “ピッチアクセントが適切でない単語の知覚：fMRIによる観測,” 日本音響学会秋季大会予稿集 2-Q-10, 2007.

[14]勝瀬郁代,“調波性の程度を考慮したPHONOBEST認識実験,” 電気関係学会九州支部連合大会予稿集, 2007.

[15]H. Kawahara, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” Proc. of Eurospeech99, pp.2781－2784, 1999.

[16]P. Lockwood and J. Boudy,“Experiments with a Non-linear spectral subtractor (NSS), hidden Markov models and the prediction, for robust speech recognition in cars,” Proc. of Eurospeech1991, pp.79－82, 1991.

[17]産業技術総合研究所RIO-DB単語音声データベース http://riodb.ibase.aist.go.jp/db066/abst.html