仮説生成／選択による大語彙連続音声認識アルゴリズムの統一的表現とそれに基づく汎用デコーダの開発

学籍番号　２３４１３５４８　氏名　土屋 貴裕
指導教員名　李 晃伸

1　はじめに

近年，音声認識技術は幅広い分野で利用されている．特に数万語彙以上の大規模なタスクにおいて発話内容をリアルタイムに認識する大語彙連続音声認識では，音声の特徴や単語の並びを確率統計的な手法で学習したモデルを利用している．それらはそれぞれ音響モデル，言語モデルと呼ばれ，モデルを利用して認識を行う音声認識の中核となるシステムを音声認識デコーダと呼ぶ．

音声認識デコーダは様々な改良により実装やアルゴリズムが state-of-the-art と呼ばれるまでに発展し，その結果，効率良く高精度な認識が可能になっている．しかし，その結果多くのデコーダにはアドホックでヒューリスティックな手法が多く導入されてきており，全てのアルゴリズムが理論的に明かになっておらず，それらが複雑に絡み合うことでデコーダ全体を把握することは非常に難しくなっている．しかし，それらの手法が効率的で高精度な認識に繋がっているのは事実であり，それらを統一的な枠組みで表現し，関連付けてより簡潔な仕組みとして体系立てていくことが今後の音声認識の発展に繋がると考えられる．

近年では重み付き有限状態トランスデューサ (WFST) [1] に基づくデコーダが注目を集めている．WFST は従来の複雑な仕組みを持つデコーダに対して，モデルに対して汎用性が高く探索の仕組みが簡明な枠組みである．しかし，現在の WFST 型デコーダでは従来のデコーダで用いられてきた言語モデル・音響モデル・発音辞書といった階層ごとの性質や階層間の関係を利用あるいは入力照合と並行した制御といった各種技法を直接的に表現することはできず，性能の面で最適とは言えない．

そこで本稿では統一的な枠組みで従来のアルゴリズム・探索技法を統一的に表現することを第一の目標とした定式化を行う．そして，その成果を新たな汎用音声認識デコーダとして実装する．統一的な枠組みを実現するために WFST を採用し，この上で様々なアルゴリズムを表現するための仕組みについて提案する．実装するデコーダは様々な従来の技法を表現することが可能な柔軟かつ汎用的なデコーダであり，アルゴリズムを自由に組み替えることで自在に音声認識システムを構築することを目指すものである．

2　デコーダ

デコーダは入力特徴量に基づいて単語列を出力する音声認識の中核となるシステムである．デコーダは音響モデルと言語モデルの確率の積が最大となるような単語列を探索する．特に大語彙では考慮すべき仮説数が増大するため，各種モデルをいかに効果的かつ効率的に利用して最適な解を求めるかというアルゴリズムが重要になる．

連続音声認識における代表的な探索アルゴリズムとして，フレーム同期ビーム探索と呼ばれるアルゴリズムが存在する．木構造化辞書と呼ばれる全単語候補を効率良くまとめて表現しながら，入力された音声をフレームごとに同期して尤度計算をしつつ仮説を順次展開することで認識を 1 パスで効率良く行う．毎フレームの仮説展開後に，スコアが低い系列は最終的な仮説の一部になる見込みが薄いと判断し，スコアの上位から一定数（ビーム幅）のノードを計算する枝刈りと呼ばれる手法で仮説の削減を行う．大語彙においては非常に広いビーム幅で探索を進める必要があるため．段階的な仮説の絞り込みを行って少ないビーム幅で効率的に探索を行うマルチパスでコーディングが用いられる．

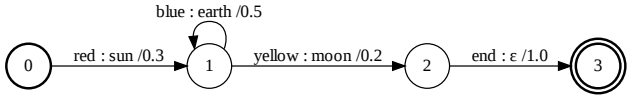

図 1: 重み付き有限状態トランスデューサ

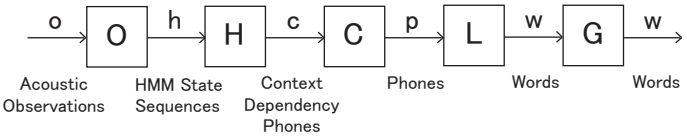

図 2: 各階層の WFST（合成前）

3　WFST に基づく音声認識

WFST は有限オートマトンを拡張した形で定義される．WFST の例を図 1 に示す．WFST は遷移に入力シンボル・出力シンボル・重みを持つオートマトンである（図中では入力：出力 / 重みとして記述）．この WFST に対して「red, blue, yellow, end」というシンボル系列を入力すると「sun, earth, moon」が出力される．ε は空語を表すため，シンボルとしては何も出力がされない．音声認識では重みは対数値として扱われることが多いため，前述した入力シンボル系列に対する重みは 2.0（0.3+0.5+0.2+1.0=2.0）となる．

WFST は数学的に簡明な枠組みであり，合成・決定化・最小化などの演算を事前に適用してネットワークの最適化を行うことができる．音声認識で特に重要となるのは合成演算であり，2 つの WFST を入力とし，新しく 1 つの WFST を出力する．合成後の WFST である $T_1 \circ T_2$ は入力として T_1 の入力を持ち，出力として T_2 の出力を持つ WFST となる．WFST は様々な形式のモデルを WFST の形に落とし込むことで全てのモデルを統一的に扱うことが可能で，オートマトンという単純な構造を持つためモデルに対する汎用性が高い．

WFST に基づく音声認識では，各種モデルを個々の WFST として構築する．そして，それらを合成することで単一のオートマトンで表現し，その上で探索を行う．図 2 に合成前の各

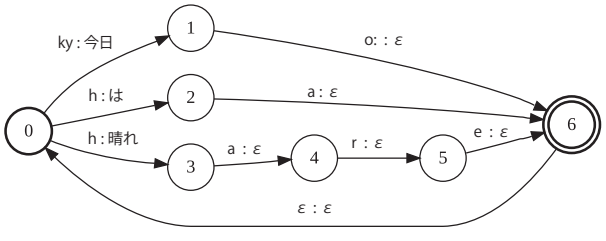

図 3: 発音辞書の例

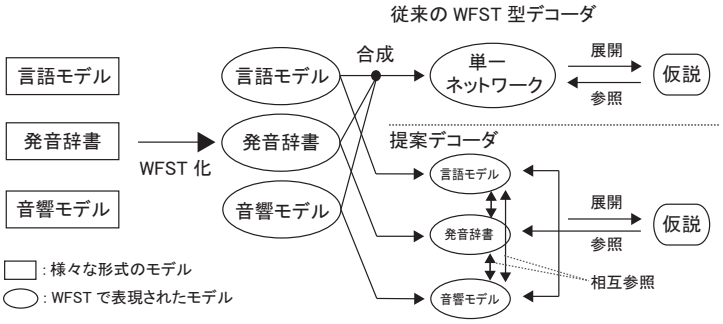

図 4: 従来の WFST 型デコーダと提案デコーダの枠組み

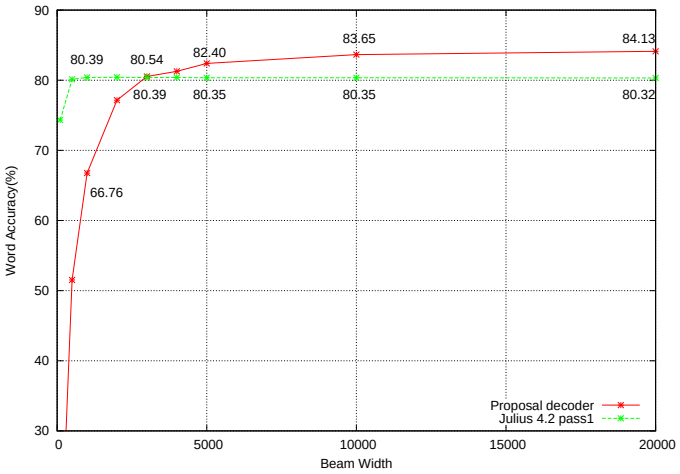

図 5: 実験結果

階層の例を示す．O, H は HMM である音響モデルから生成され，それぞれ出力確率と遷移確率を重みとして持つ．C は音素コンテキスト情報から生成され，モノフォン音素をトライフォン等のコンテキスト依存音素へと変換する．L は発音辞書から生成され，音素を単語へと変換する（図 3）．G は言語モデルから生成され，言語重みを付与する．合成後の WFST は $O \circ H \circ C \circ L \circ G$ として表現され，合成前の O の入力と G の出力を持つ．合成後は特徴量を入力として単語列を出力する単一の WFST となる．

4　仮説生成／選択によるアルゴリズムの統一的表現とそれに基づく汎用デコーダ

本稿で提案するデコーダは，統一的な枠組みを実現するための枠組みとして WFST を採用する．従来の WFST 型デコーダでは単一の WFST を参照して探索を行うのに対して，提案デコーダでは複数の WFST を相互に参照して探索できる仕組みを持つ．従来の WFST 型デコーダの枠組みと提案デコーダの枠組みの対比図を図 4 に示す．上側が従来の WFST 型デコーダ，下側が提案デコーダの枠組みである．

従来の WFST の枠組みでは様々な形式のモデルを WFST という枠組みに落とし込んだ上で，それらを単一のオートマトンネットワークに合成して探索を行う．探索時は WFST を最適化した上でどのような探索を行うかの指標は与えず，あくまで数学的な枠組みの上で探索を行う．単一のネットワークへ合成した段階で，各階層情報が消えるため，従来のデコーダが効率良く探索を行うために用いている階層ごとの性質や階層間の関係が利用できなくなる．そのような理由で，現在の WFST 型デコーダは最適な探索が行える枠組みとは言えない．そこで，それら階層情報を細かく参照できる仕組みをデコーダに持たせることで従来のデコーディングアルゴリズムが実現できる．

本稿で提案するデコーダは，それらの階層情報を効果的に利用するために，探索空間と仮説空間を階層化したまま探索が行える仕組みを提供する．探索空間とは全ての仮説を含む空間すなわちモデルのことである．対する仮説空間は探索空間上の探索中に辿ったパスのみを含む空間で，探索空間の一部である．本デコーダでは各空間はどちらも WFST で表現され，仮説空間は各階層の状態を結ぶリンクを持つ．これにより一般的な WFST 型デコーダの持つモデルに対する汎用性を損なわず，階層構造を維持したまま仮説の展開を行うことで現在の WFST 型デコーダで扱われているアルゴリズムはもちろんのこと，従来のデコーダで行われているヒューリスティックな探索を実現でき，特定の階層の仮説を探索空間として再利用することによりマルチパスデコーディングすらも実現できる．

提案デコーダでは，よりシステマティックなデコーディングの枠組みを実現するために，探索を仮説ジェネレータ・仮説セレクタという 2 つのモジュールに分けて実現した．仮説ジェネレータは従来のデコーダにおける仮説の展開・照合が実現され，仮説セレクタでは，従来のデコーダにおける仮説の選択が実現される．仮説の枝刈りや単語履歴のマージ等はセレクタにて行われる．デコーダは特徴量の入力に対してジェネレータとセレクタを交互に駆動することにより探索を進める．

5　評価実験

大語彙タスクにおける認識性能を確認するために，評価セットとして JNAS IPA-98-TestSet を用いた評価実験を行った．語彙数は 2 万語で，評価文章数は男女計 200 文章である．

本実験では大語彙連続音声認識ソフトウェア Julius [2] のアルゴリズムである木構造化辞書と 1-best 単語履歴近似を用いたフレーム同期ビーム探索を実装して評価を行った．全てのアルゴリズムを完全に一致させて評価するのは難しいため，細かな近似アルゴリズムは実装していない．実験結果を図 5 に示す．縦軸が認識精度を表し，横軸が枝刈り時のビーム幅を表す．

結果，最高認識精度は提案デコーダが 84.13%と Julius の 80.39%を大きく上回る結果となった．ビーム幅 20k 程度でも上昇の余地を残しており，30k〜40k 程度でも若干の改善が見られた．提案デコーダでは認識精度が収束するまでに広いビーム幅が必要になるが，ビーム幅による認識精度の違いは近似アルゴリズムの違いによるものであるため，本実験によって大語彙において十分な認識性能が得られたと言える．

6　むすび

本稿では，統一的な枠組みで従来のアルゴリズム・探索技法を統一的に表現することを第一の目標とした定式化を行い，その成果を新たな汎用音声認識デコーダとして実装した．

今後の課題として，様々なアルゴリズムを提案デコーダで実装・比較を行うこと，デコーダの高速化・省メモリ化が挙げられる．

参考文献

[1] M. Mohri, *et al.*, “Weighted Finite-State Transducer in Speech Recognition,” in Proc. of Computer Speech and Language, vol.16, no.1, pp. 69-88, 2002.

[2] A. Lee, *et al.*, “Recent Development of Open-Source Speech Recognition Engine Julius,” in Proc. of APSIPA, pp. 131-137, 2009.