音声言語情報処理 14－14
(1996. 12. 13)

単語 trigram を用いた大語彙連続音声認識

吉田航太郎* 松岡達雄** 大附克年** 古井貞熙* **

*東京工業大学情報理工学研究科
〒152 東京都目黒区大岡山 2-12-1
**NTT ヒューマンインターフェース研究所
〒180 東京都武蔵野市緑町 3-9-11

あらまし 大語彙連続音声認識システムの性能は、音響モデルだけではなく、使用する言語モデルの性能にも大きく依存する。本研究では、言語モデルとして日本語で初めて、マルチパスアプローチによって単語 trigram を適用した。語彙を 7000 語に限定した上で、日経新聞約 5 年分のテキストから言語モデルを学習し、不特定話者の音声を用いた新聞文章の読み上げタスクによって評価を行った結果、従来の単語 bigram 言語モデルを用いた場合より誤り率が約 44%削減され、単語正解精度で約 90%の性能を得ることが出来た。

キーワード 連続音声認識、大語彙、言語モデル、単語 trigram

Large-Vocabulary Continuous Speech Recognition Using Word Trigrams

Kotaro Yoshida* , Tatsuo Matsuoka** , Katsutoshi Ohtsuki** , and Sadaoki Furui* **

* Tokyo Institute of Technology
2-12-1 Ookayama Meguro-ku, Tokyo 152
** NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180

Abstract Accuracy of Large-Vocabulary Continuous Speech Recognition(LVCSR) systems depend not only on the performance of its acoustic models but also on the performance of the language models. We successfully incorporated word trigram language models to a LVCSR system for the first time in Japanese speech recognition. For a 7000-word vocabulary speaker-independent recognition task, trigram language models were trained using texts from a CD-ROM compilation of a Japanese news paper(Nikkei 1990-1994). Evaluation tests show that 44% reduction in word error rate from our previous word bigram system was achieved resulting in 90% word accuracy.

Keywords large vocabulary continuous speech recognition, word trigram, language models

1.はじめに

大語彙連続音声認識は、ARPA(Advanced Research Projects Agency)のWSJ-NAB(Wall Street Journal-North American Business News)コーパスなどを用いて多くの研究機関で盛んに研究されている。アメリカ英語の他にも、イギリス英語、フランス語、ドイツ語、イタリア語などを対象に各国の研究機関が新聞記事読み上げコーパスを用いて大語彙連続音声認識の研究を進めている[1-9]。日本語に関しても、情報処理学会と音響学会が共同して、研究機関を越えて共有できるデータベースの構築が進められており[10]、WSJ-NABタスクと同様の大語彙連続音声認識の研究が本格化の兆しを見せている。

我々は、日本語を対象とした大語彙連続音声認識の研究を進めるため、WSJ-NABと同様の新聞記事読み上げコーパスを設計し、大語彙連続音声認識研究用データベースを構築した。形態素解析を用いることにより、通常単語分かち書きされることのない日本語においても、単語n-gramの推定を可能とした。言語モデルとして、約5年分の新聞記事から推定した単語bigramを適用した大語彙連続音声認識実験を行い、日本語に対しても統計的言語モデルが非常に有効であることを示した[11-15]。

単語n-gram言語モデルは、nを大きくすることでより長いコンテキストを考慮した精度の高いモデル化が可能であるが、信頼性の高いモデルを推定するには非常に大規模なテキストコーパスが必要となるうえ、認識システムに実装するためにも効率のよい探索を行うための工夫が不可欠である。海外では、CMUで作成された単語trigram言語モデルが、ARPAに参画している研究機関を中心に共通利用されており、多くの研究機関で単語trigram言語モデルを実装した高精度の大語彙連続音声認識システムが実現されているが[16-28]、日本語の大語彙連続音声認識システムでは単語trigram・単語4-gramが実装された例はこれまで報告がない。

我々は、単語bigramが大語彙連続音声認識システムの性能向上に非常に有効であったことから、単語trigram・単語4-gramなど、より高精度な統計的言語モデルにより、さらに性能向上が期待できるであろうと考えた。本報告では、単語trigram・単語4-gramを大語彙連続音声認識システムに適用するためのマルチパス探索の方法と、それを用いた大語彙連続音声認識実験において認識性能向上が得られることを述べる。

2.マルチパス探索による高精度言語モデルの導入

大語彙連続音声認識は、入力音声に対して尤度最大となるパスをネットワーク内で発見する問題ととらえることができる。ネットワークを構成する音響モデル/言語モデルを細密化し、詳細にすることで、より精度の高い認識を行えると期待できるが、ネットワークが複雑で膨大なものとなってしまい、探索が非常にコストのかかるものとなってしまう。そこで、探索を複数の段階に分けて行い、まず、最初に比較的精度の低い簡単なモデルを用いた探索を行い、候補を絞り込んだ後、より精度の高いモデルを用いて再度絞り込まれた候補に対して探索を行う方法が、効率的に高精度なモデルを用いる探索方法として有効である[29]。

我々は、探索を2段階に分け、最初の探索(first-pass探索)では、言語モデルとして単語bigramを用いてN-best候補を出力し、次にこのN-best候補に対して単語trigramを適用して尤度の再評価を行い(second-pass探

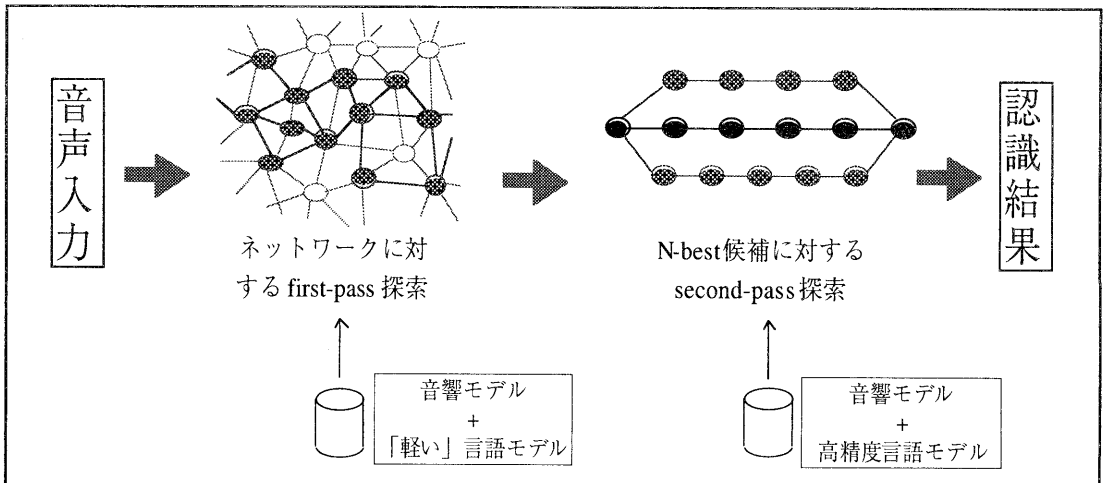

図1　**2-pass**探索による音声認識システム

索)、認識結果を得た（図 1）。今回の実験では、使用する音響モデルは、first-pass、second-pass とも同じ単語内音素文脈依存モデル（モデル数748）とした。

first-pass 探索で複数の候補を残す方法としては、尤度順で上位 N 個の候補文を生成させる N-best 形式と、候補文を各パスとして保持し、それ自身が文法ネットワークであるような word lattice 形式が考えられる。N-best の場合は当然 N を大きくすることによってより多くの候補文を残すことができ、候補文中で最も正解に近い文の単語正解精度(累積単語正解精度)を高くすることができる。一方、first-pass 探索で多くの候補文を残せば、second-pass 探索に於ける探索空間が広がり、候補の絞り込みという本来の first-pass 探索の目的と反する結果を引き起こす。従って、十分な累積単語正解精度を確保しつつ、可能な限り小さな N で N-best 候補を求めるのが望ましい。word lattice は N を非常に大きく取った場合の N-best 候補を効率的に圧縮したものとみなすことができ、first-pass 探索の出力として N が小さい N-best 候補で十分な場合、word lattice を採用する必然性はなく、逆に second-pass 探索の探索空間を広げてしまうため望ましくない。

3.実験条件

3.1.使用コーパス[11]

テキストデータベースとしては、1990 年から 1994 年までの日本経済新聞の記事 5 年分、約 750 万文を用いた。5 年分のうち、1990 年 1 月から 1994 年 9 月までの 57 カ月分(95%)を言語モデルの学習に用い(学習セット)、1994 年 10 月から 1994 年 12 月までの 3 カ月分(5%)を評価用のデータとして用いる(テストセット)。

単語 n-gram 言語モデルの学習には、学習用のテキストが単語単位で分かち書きされている必要があるが、通常、日本語のテキストは明示的な単語区切りを持っていない。そこで、テキストデータは形態素解析ツールによって形態素に分割する。以後特に断らない限り、本研究では「形態素」を「単語」として扱う。また、オリジナルのテキスト中には読み上げ不可能な記号や注釈を目的とした括弧などが含まれているが、これらを適当に削除したり、極端に長い文や形態素解析誤りを起こした文を除くなどの前処理を施した後にテキストデータベースを構築している。前処理の詳細は文献[11]を参照されたい。

実験は、語彙を出現頻度の高い上位 7000 語に限定して行った。テキストデータベース全体に含まれる語彙種類数は約 60 万語であるが、上位 7000 語彙によって全テキストの 90.3%をカバーしている。これは、WSJ の 5000 語彙限定タスク(カバー率 91.7%)に対応するものである。

音声データベースは、上記の 7000 語彙のみからなる文をテキストデータベース中から選択し、複数の文が複数の話者によって発声されている。本稿で述べる実験では、10 人の男性話者がそれぞれ互いにユニークな各 10 文をテストセットテキスト中から発話したデータ(全 100 文)を評価用に用いる。

3.2.言語モデル

言語モデルとして、上記テキストデータベースの学習セットを用いて、単語 unigram、単語 bigram、単語 trigram、単語 4-gram をそれぞれ学習した。学習セット中に観測された各言語モデルの種類数と、各言語モデルの test set perplexity は表 1 の通りである。この表からも分かる通り、n=2(bigram)以上の n-gram では、観測された種類数は、全組み合わせ種類中のごく一部である。例えば、7000 語彙の bigram には全組み合わせにして $7k^2$=49M 種類が考えられるが、実際に学習データ中に観測されたのは 2.2M 種類と、全体の 5%にも満たない。従って、n=2 以上の n-gram 言語モデルに対してはスムージングが必要となる。我々は、Katz の back-off スムージング[30]を用いた。

表1　各言語モデルの種類数と test set perplexity

	unigram	bigram	trigram	4-gram
観測された種類数	7k	2.2M	17.6M	43.3M
test set perplexity	474	48	24	18

3.3.音響モデル[11]

音響モデルには、無音を含めて 42 種類の音素から成り、音素文脈独立モデル、単語内音素文脈依存モデル(diphone/triphone context)を組み合わせた、全 748 種の sub-word 音響モデルを用いる。この音響モデルは、ATR B セットによって初期モデルを学習し、日本音響学会連続音声データベースの中の 503 文と模擬対話を用いて連結学習を行ったものである。合計で 58 名の男性話者による 15000 文を学習に用いた。

各音素 HMM の状態数は 3 で、各状態は 4 混合の混合ガウス分布とした。音声のサンプリングレートは 12KHz、量子化は 16bit である。特徴量には、フレーム長 32ms、フレームシフト 8ms で LPC 分析した 16 次の LPC ケプストラムと正規化対数パワー、及び、それらの一次時間微分を用いている。

3.4.評価方法

認識結果の評価には、認識結果と正解単語列の単語単位の DP マッチング結果に基づいて次の式で与えられる単語正解精度(Accuracy)と単語正解率(%Correct)、

単語誤り率(%Error)を用いる[31]。各単語について、正解と認識結果の文字がすべて一致する場合のみ、正しいとした。DPマッチングを行う際、文頭・文末記号、句読点、スペースは無視している。

$$Accuracy = \frac{N-S-D-I}{N} \times 100$$

$$\%Correct = \frac{N-S-D}{N} \times 100$$

$$\%Error = 100 - Accuracy$$

N:正解文の単語数　　S:置換(Substitution)

D:脱落(Deletion)　　I:挿入(Insertion)

ただし、例えば「～かどうか」という言い回しが「か」－「どう」－「か」と形態素解析されたり「か」－「どうか」になったりと、正解と認識結果の間で、形態素の分割の仕方に揺らぎがあるケースが観測されている。また、「あがる」、「上がる」といった漢字表記の揺らぎも観測されており、音声認識の立場からすると正解とすべきこのような揺らぎが、上記の評価尺度では誤りとして含まれている。これらの現象は、英語などの言語では発生しにくい、日本語固有の問題と考えられる。

4.認識実験

4.1.first-pass 探索

first-pass 探索には、音響モデルとして 748 音素の単語内音素文脈依存モデルを、言語モデルとして単語 bigram モデルを使用した。図 2は本実験の first-pass 探索に於ける、N-best の N と累積単語正解精度の関係を示したグラフである。このグラフから N の値は 300 程度で十分であり、それ以上大きくしても累積単語正解精度の大幅な向上は望めないことがわかる。従って、我々は first-pass 探索の出力として 300-best を用いることとした。

評価テキストに対する単語 bigram・300-best 認識結果の累積単語正解精度は94.0 %である。second-pass 探索では 300-best に含まれない文を選ぶことはできないから、この値は second-pass 探索の認識結果の単語正解精度の上限値となる。

一方、2-pass のシステムによって高精度な言語モデルを組み込む際のベースライン、すなわち単語 bigram 言語モデルによる認識結果の単語正解精度は、81.9 %である。

4.2.second-pass 探索

second-pass 探索には、first-pass 探索時に得られた音響モデルのスコアと、単語 trigram 言語モデルのスコア

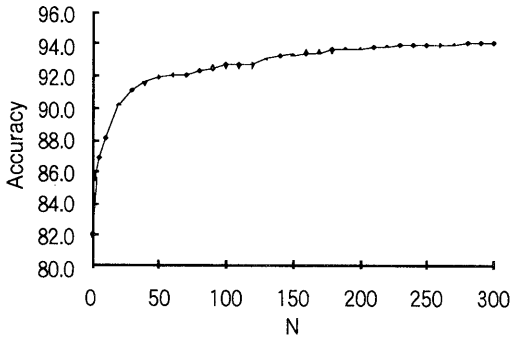

図2　単語 **bigram** を用いた **first-pass** 探索による **N-best** の累積単語正解精度

を用いている。認識性能は表2の通りである。ここで、「単語 bigram」は、first-pass 探索でも用いた単語 bigram 言語モデルによる認識性能、「300-best 累積」は first-pass 探索の単語 bigram による 300-best 累積単語正解精度を表している。

単語 bigram との性能比較を単語誤り率で行うと図 3 の通りである。単語誤り率 6.0%は first-pass 探索の累積単語正解精度から決まる second-pass 単語誤り率の下限

表2　認識性能

	単語 bigram	単語 trigram	300-best 累積
Accuracy	81.9	89.9	94.0
%Correct	83.4	90.5	94.4
%Error	18.1	10.2	6.0

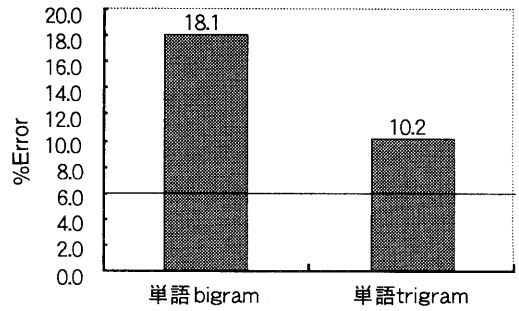

図3　単語誤り率の比較

値である。

これらの図表からわかる通り、second-pass 探索の単語 trigram によって、単語誤り率は単語 bigram によるシ

ステムに対して約44%削減され、約90%の単語正解精度を得ることができた。また、second-pass 探索の単語誤り率の下限が6.0%であることを考えると、ベースラインである18.1%から見て、second-pass で回復可能な12.1%の誤りうち、およそ3分の2にあたる7.9%が単語trigram 言語モデルによって削減されている。表1に見られるtest set perplexityの低下(48→24)から期待されたとおりの性能向上が得られたと言える。今回、時間的な制約もあり、単語4-gramを用いた認識実験は行っていないが、test set perplexityの値が18であることから、単語4-gramを用いた場合、単語trigramより更に高い認識性能が期待できる。

5. まとめ

本稿では、日本語の大語彙連続音声認識システムに初めて単語trigram言語モデルを適用し、単語bigram言語モデルを適用した場合に比べて高い認識性能が得られることを報告した。また、単語4-gram言語モデルに関しては、そのtest set perplexityの低さから、単語trigramを上回る認識性能が期待される。

今回述べたような2-pass 探索によって高精度言語モデルを組み込んだシステムでは、性能向上のための取り組みは、first-pass 探索の累積正解精度の向上と、second-pass 探索の N-best 再評価による認識性能の向上という二つの独立した課題に対して行われることになる。前者に対しては、あくまでも「軽い」モデルを用いるという立場から言語モデルをこれ以上高精度化することは難しく、音響モデルの最適化が中心課題と考えられる。実際、今回報告した実験で用いた音響モデルには、話者適応や雑音・歪みに対する適応を行っておらず、また、単語内音素文脈依存モデルのみで、単語間音素文脈依存モデルは使用していない。これらの技術との組み合わせで first-pass 探索の累積正解精度が向上する可能性は高い。一方、second-pass探索の認識性能の向上のためには、今後より精度の高い言語モデルの組み込みを検討する必要がある。

また、海外では既に20k,64kという規模の大語彙で研究されていることから、より語彙数の大きいシステムの実現も今後の課題である。現在、語彙数を 30k,150kに増やしたタスクでの認識実験を検討中である。

更に、日本語特有の幾つかの問題に対しても対策が必要である。前述した形態素解析結果や漢字表記の揺らぎに関しては、実験結果についてその数を調べたところ、誤りの約4%が、これに起因するものであった。これを除くと、単語正解精度は90.3%となる。この他にも、例えば、現在は使用語彙数を限定した単語辞書を作成する際、出現頻度順に上位一定個数の単語を漢字混じりの単語で抽出した後に、その単語が取り得る読み方を与えている。その結果、「新(あたらし)」、「朝(とも)」といった、あまり使われない読み方の辞書エントリが出来、認識誤りを引き起こしている。一つの漢字に多くの読み方が存在する日本語に於いては、読み方の頻度まで考慮に入れた辞書によって性能を向上させることができると思われる。

謝辞 形態素解析ツールを提供していただいたNTTヒューマンインターフェース研究所映像処理研究部の田中一男主幹研究員に感謝します。テキストデータ(日本経済新聞CD-ROM版,1990-1994)の使用を許諾していただいた日本経済新聞社に感謝します。また、日頃御討論いただくNTTヒューマンインターフェース研究所古井特別研究室、東工大古井研究室の皆様に感謝します。

参考文献

[1] B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. ICSLP-92, pp.899-902, Oct. 1992

[2] J. L. Gauvain, L. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large French read-speech corpus, "Proc. ICSLP-90, pp.1097-1100, Oct. 1990

[3] T. Robinson, J. Fransen, D. Pye, J.Foote, and S.Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," Proc. ICASSP-95, pp.81-84, May 1995

[4] H. J. M. Steeneken and D. A. van Leenwen, "Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: SQUALE-project," Proc. EUROSPEECH-95, pp.1271-1274, Sep. 1995

[5] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young, "The 1994 HTK Large vocabulary speech recognition system," Proc. ICASSP-95, pp.73-76, May 1995

[6] D. Pye, P. C. Woodland, and S. J. Young, "Large vocabulary multilingual speech recognition using HTK," Proc. EUROSPEECH-95, pp.181-184, Sep. 1995

[7] L. Lamel, M. Adda-Decher, and J. L. Gauvain, "Issues in large vocabulary, multilingual speech recognition," Proc. EUROSPEECH-95, pp.185-188, Sep. 1995

[8] D. B. Paul, and B. F. Necioglu, "The Lincoln large-vocabulary stack-decoder HMM CSR," Proc. ICASSP-93, pp.660-663, Apr. 1993

[9] L. Lamel, and R. De Mori, "Speech recognition of European languages," Proc. IEEE Automatic Speech Recognition Workshop, pp.51-54, Snowbird, Dec. 1995

[10] 武田, 伊藤, 松岡, 竹沢, 鹿野, “大語彙連続音声認識研究のためのテキストデータ整備, ” 情処研報, Vol.96, No.55, pp.49-54
[11] 大附, 森, 松岡, 古井, 白井, “新聞記事を用いた大語彙連続音声認識の検討, ” 信学技報, SP95-90, pp.63-68, Dec.1995
[12] 森, 大附, 松岡, 古井, 白井, “新聞読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討, ” 日本音響学会春期研究発表会, 3-8-7, pp.159-160, Mar. 1996
[13] 大附, 森, 松岡, 古井, 白井, “新聞読み上げタスクを用いた大語彙連続音声認識における音響モデルの検討, ” 日本音響学会春期研究発表会, 3-8-7, pp.159-160, Mar. 1996
[14] T. Matsuoka, K. Ohtsuki, T. Mori, S.Furui, and K.Shirai, “Large-vocabulary continuous-speech recognition using a Japanese business newspaper(Nikkei),” Proc. of ARPA Speech Recognition Workshop, Feb. 1996
[15] 松岡, 大附, 森, 古井, 白井, “新聞記事データベースを用いた大語彙連続音声認識, ”電子情報通信学会論文誌 D-II, Vol. J79-D-II, No. 12
[16] R. Rosenfeld, "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPACSR Evaluation," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 47-50, Jan. 1995
[17] L. Chase, et al., "Improvements in Language, Lexical, and Phonetic Modeling in Sphinx-II," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 60-65, Jan. 1995
[18] L. Nguyen, et al., "The 1994 BBN/BYBLOS Speech Recognition System," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 77-81, Jan. 1995
[19] V. Digalakis, et al., "Continuous Speech Dictation on ARPA's North American Business News Domain," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 88-93, Jan. 1995
[20] P. C. Woodland, et al., "The Development of The 1994 HTK Large Vocabulary Speech Recognition System," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 104-109, Jan. 1995
[21] R. Roth, et al., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 116-120, Jan. 1995
[22] L. R. Bahl, et al., "The IBM Large Vocabulary Continuous Speech Recognition System for The ARPA NAB News Task," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 121-126, Jan. 1995
[23] J. L. Gauvain, et al., "Developments in Large Vocabulary Dictation: THe LIMSI Nov94 NAB System," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 131-138, Jan. 1995
[24] M. Ostendorf, et al., "The 1994 BU NAB News Benckmark System," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 139-142, Jan. 1995
[25] D. Paul, "New Developments in the Lincoln Stack-Decoder Based Large-Vocabulary CSR System," Proc. ARPA Spoken Language Systems Technology Workshop, pp. 143-147, Jan. 1995
[26] Andrej Ljolje, Michael Riley, Donald Hindle, and Fernando Pereira, “The AT&T 60,000 Word Speech-To-Text System,” Proc. of ARPA Speech Recognition Workshop, pp. 162-165, Jan. 1995
[27] J. L. Gauvain, L. Lamel, G. Adda, and D. Matrouf, “The LIMSI 1995 Hub3 System,” Proc. of ARPA Speech Recognition Workshop, pp. 105-111, Feb. 1996
[28] A. Sankar, A.Stolcke, T.Chung, L. Neumeyer, M. Weintraub, H. Franco, and F. Beaufays, “Noise-resistant Feature Extraction and Model Training for Robust Speech Recognition,” Proc. of ARPA Speech Recognition Workshop, pp. 117-122, Feb. 1996
[29] R. Schwartz, L. Nguyen, and J. Makhoul, “Multiple-pass search strategies,” in Automatic speech and speaker recognition, ed. Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, pp.429-456, Kluwer Academic Publishers, Massachusetts, 1996.
[30] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” IEEE Trans. ASSP-35, pp.400-401, Mar. 1987
[31] F. Kubala, Y. Chow, A. Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift, “Continuous speech recognition results of the BYBLOS system on the DARPA 1000-word resource management database,” Proc. ICASSP-88, pp.291-294, May 1988