音声言語情報処理 36−13
（2001. 6. 1）

音響尤度と言語尤度を用いた音声認識結果の信頼度の算出

中川 聖一　堀部 千寿

豊橋技術科学大学 情報工学系
〒441-8580 豊橋市天伯町字雲雀ヶ丘 1-1
nakagawa@slp.tutics.tut.ac.jp

最近、多くの大語彙連続音声認識システムが開発、使用されているが、音声認識された結果には、認識誤りが含まれることが多い。

そこで、音声認識の結果どの部分が正しいか、または誤っている可能性が高いかを判別できればアプリケーションに対する悪影響を軽減することができると考えられる。このような正しい（誤っている）可能性が高いかを判別するパラメータは信頼度（Confidence Measure）とよばれ、大語彙音声認識システムや対話システムなどでの利用が考えられている。

本稿では信頼度を音響的なアプローチと言語的なアプローチからそれぞれ計算し、それぞれの結果の論理和をとることにより、正解単語の判定の精度を向上させる方法を提案する。

Confidence Measures for Speech Recognition by Using Likelihood of Acoustic Model and Language Model

Seiichi NAKAGAWA and Yoshihisa HORIBE

Department of Information and Computer Sciences
Toyohashi University of Technology, Tenpaku, Toyohashi, 441-8580, Japan
nakagawa@slp.tutics.tut.ac.jp

The recognition errors are inevitable for large vocabulary continuous speech recognition systems. If unreliable candidates are correcty identified, the harmful influence caused by recognition errors will reduce. The measure of reliability is called "Confidence Measure" and it is useful for various applications such as transicription systems and dialogue systems.

In this paper, we propose a new conefidence measure which combines logically the likelihood of acoustic model and that of language model.

1 はじめに

最近、多くの大語彙連続音声認識システムが開発、使用されているが、音声認識された結果には、認識誤りが含まれることが多い。そのため、他のアプリケーションなどでこれらの音声認識の結果を使う場合、この認識誤りがアプリケーションの動作に悪い影響を与えることが考えられる。

そこで、音声認識の結果のどの部分が正しいか、または誤っている可能性が高いかを判別できればアプリケーションに対する悪影響を軽減することができると考えられる。このような正しい（誤っている）可能性が高いかを判別するパラメータは信頼度（Confidence Measure）とよばれ、大語彙音声認識システムや対話システムなどでの利用が考えられている。信頼度には単語のグラフから事後確率を利用するものや、N-bestリスト、ワードラティスなどから算出するものなど様々な研究が行われている[1][2][3][4][5] [6]。

[1] では、従来から信頼度としてよく用いられている単語の事後確率の求め方を改良したもの（単語候補の境界を変化させる）やN-bestリストから求めるもの、N個のbestリスト中に同じ時刻に同じ単語が現れる割合を事後確率として定義したもの(Acoustic Stability[7])、word graph中でのある時刻での単語仮説の数を信頼度として定義したもの(Hypothesis Density[8])を、比較実験している。

[6] では、認識単語と音節連接モデルとの尤度比（LLR）と、単語内の音節継続時間の分散（VSD）をそれぞれ用いてシグモイド関数を定義し、それによる誤認識単語のリジェクションを行ない、LLRとVSDの組み合わせによりリジェクションの精度を向上させている。また、[9] では、複数の認識システムの認識結果を統合して、信頼度のある認識結果を出力する方法が試みられている。

本稿では信頼度を音響的なアプローチと言語的なアプローチからそれぞれ計算しそれぞれの結果の論理和をとることにより、正解単語の判定の精度を向上させる方法を提案する。

2 信頼度の定義

単語を W 、入力音声を A 、p を任意の音節列とした場合、音声認識は以下の事後確率を最大化する問題と捉えることができる。

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)} \approx \frac{P(A|W)P(W)}{\max_p P(A,p)} \quad (1)$$

この式は単語のみに依存する項と入力音声に依存する項に分割することができる。ここで、入力音声に依存する項 $\frac{P(A|W)}{\max_p P(A,p)}$ を音響レベルの信頼度、また単語のみに依存する項 $P(W)$ を言語レベルの信頼度として利用できると考えられる。以下にそれぞれの信頼度の定義を述べる。

2.1 音響尤度を使用した信頼度

はじめに、音響レベルでの信頼度の計算法を示す。(1) 式の入力音声に依存する項をもとに、p を任意の音節列とすると、音響尤度を使用した信頼度を以下のように定義する。

$$C_a^i = S_o(t_{si}, t_{ei}) - S_a(t_{si}, t_{ei}) \quad (2)$$

ここで、C_a は単語 W_i における音響的信頼度を表す。また、t_{si} と t_{ei} はそれぞれ単語 W_i の開始、終了時刻とする。$S_o(t_{si}, t_{ei})$ は時刻 t_{si} から t_{ei} での最適な音節列の（対数）音響尤度とする。$S_a(t_{si}, t_{ei})$ は単語 W_i の（対数）音響尤度である。

2.2 言語尤度を使用した信頼度

次に、言語レベルでの信頼度の計算法を示す。(1) 式の単語にのみ依存する項をもとに、言語尤度を使用した信頼度を以下のように定義する。

$$C_l^i = \log \frac{1}{P_l(W_i|W_1^{i-1})} \quad (3)$$

ここで、C_l^i は単語 W_i における言語的信頼度を表す。また、$P_l(W_i|W_1^{i-1})$ は単語 W_i の推定出現確率である。この式は言語レベルの信頼度に、単語 W_i の単語パープレキシティの対数（エントロピー）を用いることを表している。

2.3 音響尤度と言語尤度の統合法

（1）式を信頼度に用いる場合は、（2）式と、（3）式の和になる（通常は、（3）式に言語重みを乗じる）。

本稿では、音響尤度と言語尤度を独立に用いる場合、およびAND条件、OR条件を検討する。

3 認識実験

2 で定義した信頼度の有効性を確かめるために、大語彙連続音声認識システム (SPOJUS) で認識実験を行った [10]。実験条件は、表 1 に従う。本システムは、本来 2 パス方式であるが、1 パス目 (言語モデルはバイグラム) の結果を用いた。この実験条件におけるテストセット (男性 9 名による 100 文の読み上げ音声) の単語正解率 (Cor) は 84.9 %、正解精度 (Acc) は 80.9 % であり、評価データの単語数は 1588 単語であった。

表 1: 実験条件

(A) : 音響的条件

サンプリング周波数	12 kHz
窓関数	21.33 ms ハミング窓 (256 points)
フレーム周期	8 ms (96 points)
分析方法	14 次元 LPC 分析
特徴パラメータ	$LPC\ MEL\ CEP$ (10 次元 × 4 フレームを KL 展開により 20 次元に圧縮) $+\Delta CEP$ (10 次元) $+\Delta\Delta CEP$ (10 次元) $+\Delta POW + \Delta\Delta POW$ (計 42 次元)
音響モデルの単位	音節 (正確にはモーラ単位)
音節種類数	114
HMM	連続出力分布型 HMM (5 状態 4 出力分布、全共分散行列、4 混合ガウス分布)
総分布数	1824 分布
音響モデル学習データ	ASJ データベース (男性 30 名 4518 文) 新聞記事読み上げ音声 (JNAS)(男性 125 名 12703 文)

(B) : 言語的条件 (20,000 語彙の場合)

語彙数	20,000
言語モデル	単語 bigram(1 パス目 : パープレキシティ = 117.1)
言語モデル学習データ	JNAS 1991 年 1 月~1994 年 9 月 (約 330 万文)
評価データ	100 文 (JNAS より話者 9 人分)
未知語	6 単語 (未知語率 0.4%)

3.1 音響信頼度のみの場合

音響信頼度のみを使用した場合の単語の正解率と信頼度の関係を、図 1 に示す。

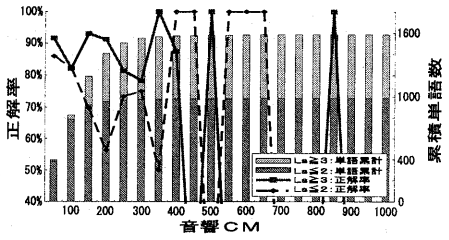

図 1: 音響信頼度と単語正解率の関係 (SPOJUS)

図 1 において、横軸は式 (2) により計算された値、左側の縦軸は単語正解率を、右側の縦軸は単語の累積度数を示す。図 1 の結果は、式 (2) を用いて音声認識の結果について音響信頼度を計算し、その音響信頼度を元にしてヒストグラムを作成し、ヒストグラムの各区間で単語認識率を計算し、プロットしたものである。また、図中では単語の音節数 Ls により短い単語 (音節数 Ls が 2 以下の単語、点線で表記) と長い単語 (音節数 Ls が 3 以上の単語、実線で表記) に分けて結果を示す。単語の長さで結果を分ける理由については、式 (2) の音響信頼度では単語の長さを考慮していないため、短い単語ほど音響信頼度が (正解率に関係なく) 良く計算されるためである。

図 1 の結果より、2 音節以下の短い単語に関しては精度良く (正解率が 90 %以上) 正解単語を判別することができず、全体として正解率が悪くなっている。一方、3 音節以上の長い単語に関しては、音響信頼度の値が 200 以下の部分で精度良く (正解率が 90 %以上) 正解単語を判別できる。この判別法では信頼度が 200 以下になった 454 単語について正解率 90 %以上で正解単語と判断でき、これは 3 音節以上の単語の 75.2 %、全単語の 28.6 %に相当する。また、正解単語と判断された単語全体の正解率は 91.6 %であった。

3.2 言語信頼度のみの場合

言語信頼度のみを使用した場合の単語の正解率と信頼度の関係を、図 2 に示す。

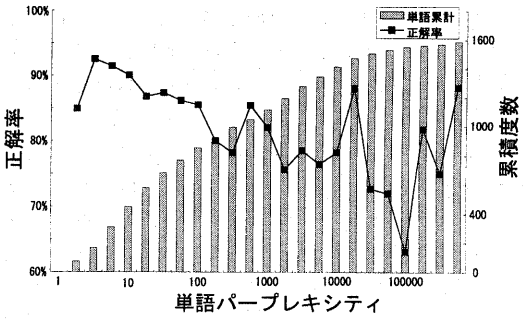

図 2: 言語信頼度と単語正解率の関係 (SPOJUS)

図 2 において、横軸は式 (3) により計算された値 (便宜上単語パープレキシティで表記)、左側の縦軸は単語正解率を、右側の縦軸は単語の累積度数を示す。図 2 の結果は図 1 と同様の方法で作成したヒストグラムである。

図 2 の結果より、言語信頼度の値 (実際はパープレキシティで説明する。以下同様) が 10 以下の場合に正解率が 90 %以上となっている。この判別法

ではパープレキシティが 10 以下になった 376 単語について正解率 90 %以上で正解単語と判断でき、これは全単語の 23.7 %に相当する。この正解単語と判断された単語全体の正解率は 91.2 %であった。

また、言語信頼度である単語パープレキシティ (対数をとるとエントロピー) に正解率がほぼ反比例するという傾向が見られた [10]。このことから、単語パープレキシティが信頼度として使用できることが実験的に確かめられた。

3.3 音響・言語信頼度を組み合わせた場合

次に音響信頼度と言語信頼度を組み合わせた場合について考える。一つ目の組み合わせ手法として、音響信頼度と言語信頼度を $C = C_a + \lambda C_l$ のように重み付き和で正解単語の判別を行った (通常の事後確率法)。λ は認識の際の言語重みと同じ値を使用した。二つ目の組み合わせ手法として、音響信頼度または言語信頼度のどちらかで正解率 90 %以上で正解単語と判定される場合は正解単語とする (OR 条件) 方法を用いた。また、音響信頼度と言語信頼度の両方で正解率 90 %以上で正解単語と判定される場合を正解単語とする (AND 条件) 方法も用いた。それぞれの結果を表 2 に示す。この表で、" 音響 CM " は音響信頼度、" 言語 CM " は言語信頼度を表す。

表 2 の結果より、重み付き和で正解単語の判定を行った場合、音響信頼度や言語信頼度を単独で使用した場合と比べ、精度良く判定できる単語が 2 %程度増えたが、あまり大きな改善とはいえない。

しかし、OR 条件を用いると正解率 90 %以上で正解単語と判定される単語数は 772 単語、全単語の 48.6 %であり、音響信頼度や言語信頼度を単独で判定する場合や、重み付き和で判定する場合と比べ、大きく改善されている。

表 2: 音響信頼度と言語信頼度を組み合わせた場合

	正解率が 90 %以上で判定できる単語数	割合 [%]	正解率 [%]
音響 CM のみ	454 単語	28.6	91.6
言語 CM のみ	376 単語	23.7	91.2
重み付き和	486 単語	30.6	92.2
OR 条件	772 単語	48.6	90.9
AND 条件	109 単語	6.9	98.1
	123 単語	8.0	95.9

また、" 音響 CM のみ " で正解と判定され、かつ " 言語 CM のみ " でも正解と判定される単語 (AND 条件) の数はしきい値動作によって異なるが一般に高い信頼度で正解を判定できるが、判定できる単語数が 100 単語前後 とかなり少数であることから、音響信頼度のみで正解と判定される単語の集合と言語信頼度のみで正解と判定される単語の集合はほぼ独立であるといえる。

このことから、信頼度は音響的な信頼度と言語的な信頼度を分けて計算し、OR 条件で正解かどうかを判別する手法が有効であると言える。

以下の図 3 に重み付き和により正解と判定された単語の例を、また、図 4 に OR 条件により正解と判定された単語の例を示す。ゴシック体が信頼度の高い単語と判定されたものである。なお、各単語の添字の C は正解、I は挿入誤り、S は置換誤りを示している。正解入力文は以下の通りである。

- お 年寄り から の 注文 に も 備え 二千 個 分 の 材料 を 確保
- 個人 技 が 随所 で 光っ た
- 警察 が 摘発 し た 汚職 事件 の わいろ 額 で は 過去 最高
- 下 の 弟 の 声 が 聞こえ た
- そして 暫定 的 措置 と して 一人 当たり 四万 ドル が 提示 さ れ て いる

- 年寄り$_C$ から$_C$ の$_C$ 中$_I$ の$_S$ に$_C$ も$_C$ 備え$_C$ **て**$_S$ 千$_S$ 株$_S$ の$_C$ **材料**$_C$ 加工$_S$
- **個人**$_C$ 技$_C$ が$_C$ **重傷**$_S$ で$_C$ 光っ$_C$ た$_C$
- **警察**$_C$ が$_C$ 摘発$_C$ し$_C$ た$_C$ **汚職**$_C$ **事件**$_C$ の$_C$ は$_S$ 医学$_S$ で$_C$ は$_C$ 過去$_C$ **最高**$_C$
- 下$_C$ の$_C$ **弟**$_C$ の$_C$ 声$_C$ が$_C$ **聞こえ**$_C$ た$_C$
- **そして**$_C$ 暫定$_C$ 的$_C$ 措置$_C$ と$_I$ し$_I$ て$_S$ **一人**$_C$ 当たり$_C$ **四万**$_C$ ドル$_C$ が$_C$ **提示**$_C$ さ$_C$ **れ**$_C$ **て**$_C$ いる$_C$

図 3: 重み付き和により正解と判定された単語の例

- 年寄り$_C$ から$_C$ の$_C$ 中$_I$ の$_S$ に$_C$ も$_C$ 備え$_C$ て$_S$ 千$_S$ 株$_S$ の$_C$ 材料$_C$ 加工$_S$
- 個人$_C$ 技$_C$ が$_C$ 重傷$_S$ で$_C$ 光つ$_C$ た$_C$
- 警察$_C$ が$_C$ 摘発$_C$ し$_C$ た$_C$ 汚職$_C$ 事件$_C$ の$_C$ は$_S$ 医学$_S$ で$_C$ は$_C$ 過去$_C$ 最高$_C$
- 下$_C$ の$_C$ 弟$_C$ の$_C$ 声$_C$ が$_C$ 聞こえ$_C$ た$_C$
- そして$_C$ 暫定$_C$ 的$_C$ 措置$_C$ と$_I$ し$_I$ て$_S$ 一人$_C$ 当たり$_C$ 四万$_C$ ドル$_C$ が$_C$ 提示$_C$ さ$_C$ れ$_C$ て$_C$ いる$_C$

図 4: OR 条件により正解と判定された単語の例

4 他の認識システムにおける信頼度の効果

前節で行った実験と同様の実験を、本研究で標準として使用している認識システムとは異なるシステムを用いて行なった場合の結果を示す。

使用した認識システムは Julius [11] である。評価に用いたテストセットは前節で行った実験と同様のものを使用している。この実験条件におけるテストセットでの音声認識システムの認識性能は、言語モデルとしてトライグラムを用いた場合（パープレキシティは 37.7）は単語正解率（Cor）が 93.3 % であり、評価データの単語数は 1608 単語であった。評価データの単語数が SPOJUS の場合と異なるのは、形態素切りを行なうツールが SPOJUS で使用しているものと、Julius で使用しているもので異なるため（認識辞書の単語の構成単位に差が出る）である。また、Julius の音響モデルの最小単位は音素であるため、そのままでは音響 CM の計算時に必要な最適な音節列のスコア $S_o(t_{si}, t_{ei})$ の計算ができない。そこで、Julius に、音節のみを登録した単語辞書を使用することで音節認識を行なっている。この認識結果のスコア情報を用いて音響 CM を計算している。

音響 CM のみを用いた場合（図 5 参照）は、4 音素以下の場合は精度 97 %で正解単語を判別できる単語数は 11 単語（音響 CM = 5 未満）、5 音素以上では 287 単語（音響 CM = 25 未満）であった。これは全単語の 18.6 %にあたる。

言語 CM のみを用いた場合（図 6 参照）は、精度 97 %で正解単語を判別できる単語数は 379 単語

表 3: 音響信頼度と言語信頼度を組み合わせた場合 (Julius 使用)

	正解率が 97 %以上で判定できる単語数	割合 [%]	正解率 [%]
音響 CM のみ	298 単語	18.6	97.2
言語 CM のみ	379 単語	23.7	97.1
重み付き和	545 単語	34.0	97.4
OR 条件	667 単語	41.6	97.0
AND 条件	89 単語	5.5	98.9

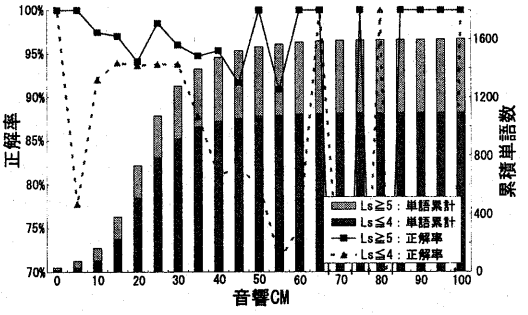

図 5: 音響信頼度と単語正解率の関係（Julius 使用）

（言語 CM = 1.25 未満、単語パープレキシティ = 2.4 未満）であった。これは全単語の 23.7 %にあたる。

音響 CM と言語 CM の重み付き和を用いた場合は、4 音素以下の場合は精度 97 %で正解単語を判別できる単語数は 504 単語、5 音素以上では 41 単語であった。これは全単語の 34.0 %にあたる。

音響 CM と言語 CM の OR 条件を用いた場合は、97 %以上の正解率で正解と判定できる単語が

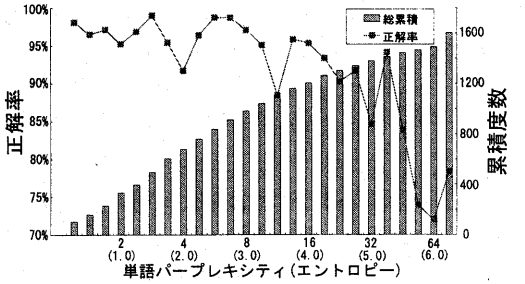

図 6: 言語信頼度と単語正解率の関係（Julius 使用）

表 4: 音響信頼度と言語信頼度を組み合わせた場合 (Julius 使用)

	正解率が 97 %以上で判定できる単語数	割合 [%]	正解率 [%]
音響 CM のみ	298 単語	18.6	97.2
言語 CM のみ	379 単語	23.7	97.1
重み付き和	545 単語	34.0	97.4
OR 条件	667 単語	41.6	97.0
AND 条件	10 単語	0.6	90.0

音響 CM のみでは 298 単語 (4 音素以下の単語の音響 CM = 5 未満、5 音素以上の単語の音響 CM = 25 未満)、言語 CM のみでは 379 単語 (言語 CM = 1.25 未満、単語パープレキシティ = 2.4 未満)、重み付き和では 545 単語、OR 条件では 667 単語であり、OR 条件のときでに一番多く正解単語を精度良く判定できる。したがって、本研究室以外の音声認識システムにおいても、信頼度は音響的な信頼度と言語的な信頼度を分けて計算し、OR 条件で正解かどうかを判別する手法が有効であると言える。

しかし、我々の研究室標準の認識システムを用いた場合と比べて、OR 条件と重み付き和の結果の差が小さくなる傾向が見られた。

また、言語信頼度を固定して (判定基準を多少緩和、言語 CM = 2.75、単語パープレキシティ = 6.7) AND 条件を適応した場合、精度良く正解単語と判別できる部分は、音響 CM が 15 以下で言語 CM (パープレキシティ) が 6.7 以下の場合の 89 単語 (全体の 5.5 %程度) である。この時の正解率は 98.9 %である。このように、判定される単語は少数だが、さらに精度良く正解単語を判定できた。

5 むすび

本稿で提案した信頼度を用いることで、正解単語の約半数に対しては、もとの全単語の平均単語と解率を向上させることができた。この信頼度を応用できるタスクとしては、以下のようなものが考えられる。

- **音声対話への応用** 信頼度の良い単語を重視して対話処理、意味理解などを行なう。信頼度を用いることによって、従来音声対話で問題となっていた、誤認識による対話処理への悪影響を軽減できる。
- **教師なし学習への応用** 信頼度の良い単語のみを使用して音響モデルなどの学習を行なう。信頼度を用いることにより、発声条件が悪い単語などを学習データから除外できるようになるため、精度の良いモデルが学習できると考えられる。

参考文献

[1] F. Wessel et al,: "Confidence measures for large vocabulary continuous speech recognition" ,IEEE Trans. Speech and Audio Process, Vol.9, No.3, pp.288–298(2001)

[2] B.Rueber,: "Obtaining confidence measures for spontaneous speech recognition", Proc. EuroSpeech , pp.739–742(1997)

[3] G.Williams and S.Renals, : "Confidence measure from local posteriori probability estimates", Computer Speech and Language, Vol.13 , No.4, pp.395–411(1999)

[4] 緒方 淳, 有木 康雄,: 「音声認識精度向上のための信頼性尺度の比較」, 電子情報通信学会技術研究報告,SP2000-94(2000)

[5] 堀 智織, 古井 貞熙,: 「信頼尺度を用いた音声自動要約の改善」, 音響学会講演論文集, 3-5-5(2000)

[6] 北岡 教英, 赤堀 一郎, 中川 聖一: 「認識結果の正解確率に基づく信頼度とリジェクション」, 電子情報通信学会論文誌, D-II, Vol.J83-D-II, No.11, pp.2160–2170(2000)

[7] "Technical Report. Interactive Systems Labs" ,ILKD (1996)

[8] T.Kemp and T.Schaaf,: "Estimating confidence using word lattice", Proc. EuroSpeech , pp.827–830(1997)

[9] 小玉 康宏, 宇津呂 武仁, 西崎 博光, 中川 聖一 :「複数の音声認識システムの出力の共通部分を用いた認識誤り検出」, 言語処理学会 第 7 回年次大会発表論文集, pp.389–392(2001)

[10] 赤松 裕隆, 甲斐 充彦, 中川 聖一 :「新聞・ニュース文の第語彙連続音声認識」, 情報処理学会, 音声言語処理研究会, SLP 21-11(1998)

[11] 中川 聖一, 伊田 政樹:「連続音声認識のタスクの複雑さを表す新しい尺度」, 電子情報通信学会論文誌, Vol.81-D II, No.7, 1491–1150(1998)

[12] 河原 達也 他:「日本語ディクテーション基本ソフトウェア (99 年度番)」, 音声言語処理学会, SLP 21-11(1998)