論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声中の単語認識における人を基準としたデコーダの性能評価
Title(English)	
著者(和文)	篠崎 隆宏, 古井 貞熙
Authors(English)	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会 2002年秋季講演論文集, Vol. , No. 2-9-13, pp. 87-88
Citation(English)	, Vol. , No. 2-9-13, pp. 87-88
発行日 / Pub. date	2002, 9

話し言葉音声中の単語認識における人を基準としたデコーダの性能評価 *

篠崎 隆宏　○古井 貞熙　(東工大)

1　はじめに

話し言葉を対象とした音声認識システムの性能は十分ではなく、実用化にはより一層の改良が必要である。話し言葉音声中には比較的認識が容易と考えられる部分や正しく認識するためには話の流れや話者に関する知識が必要となる部分、人が聞いても認識が困難である部分などが含まれているため、認識システムの改善点の候補としては、比較的単純なものから長いコンテキストを用いた高度な推論が必要と考えられるものまで、様々あげられる。しかし実際にどのような改善がどの程度必要とされているのかは、あまり分かっていない。そこで比較的対応が容易と考えられる部分における改善の可能性を探ることを目的として、数単語程度に制限されたコンテキストのみが与えられた条件のもとで、音声認識システムの認識性能を人間の単語了解度を基準として評価・分析し、問題点の検討を行った。

2　分析方法

2.1　認識タスク

認識タスクは3単語分の音声を聞き、中心の単語を与えられた語彙中から選択することとし、前後の単語は既知として与えた。デコーダと被験者に与えられる課題は同一で、表 1 に示す日本語話し言葉コーパス CSJ 中の、男性 7 人の学会講演音声から切り出したサンプルを用いた。単語の選択肢に関しても、デコーダと被験者で同一とした。語彙は CSJ 中の男性話者による 455 学会講演中に出現した上位 2 万 5 千語とした。

使用したテストセットの概要を表 1 に示す。音声サンプルの切り出しは、形態素解析器 JTAG を用いて、単語単位に区切った正解文を用いた強制アライメントにより行った。この際アライメント精度への影響を避けるため、読み付与が著しく誤っている部分は除いた。得られた音声サンプルからランダムに 500 個を抽出し、評価用サンプルとした。正解単語のうち、語彙中に存在しないものが延べ約 1%存在する。認識率の集計に際しては、人手により結果をチェックし、単純な表記の揺れによる違いは正解とした。

2.2　デコーダを用いた認識実験

デコーダによる音声認識は図 1 に示すような単語ネットワークを用い、中心の単語を選択することにより行った。言語尤度として式（1）に示す値を各候補単語に割り当てた。

$$P(w_c|w_f, w_b) \tag{1}$$

$$= \frac{P(w_f) \cdot P(w_c|w_f) \cdot P(w_b|w_f, w_c)}{\sum_w P(w_f) \cdot P(w|w_f) \cdot P(w_b|w_f, w)} \tag{2}$$

ここで w_c は中心の単語、 w_f, w_b はそれぞれ前および後ろの単語である。式（1）は trigram モデルを用いて式（2）により計算することが出来る。

表 1. 音声認識タスク

Presentation ID	Conference
A01M0035	日本音響学会
A01M0007	日本音響学会
A01M0074	日本音響学会
A02M0117	国語学会
A03M0100	言語処理学会
A05M0031	音声学会
A06M0134	社会言語科学会

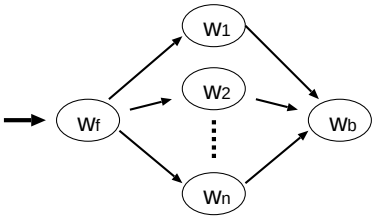

図 1. Word network for the decoder.

使用した音響モデルは、CSJ 中の男性話者による 455 学会講演の約 94 時間から学習した、2000 状態 16 混合の状態共有 triphone HMM である。音響パラメータは MFCC12 次元, Δケプストラム 12 次元, 対数パワーの 1 次差分の 25 次元で, 数秒から 20 秒程度の発話毎に, 平均ケプストラムによる正規化 (CMS) を行った。言語モデルは男性/女性話者、学会/模擬講演からなる 1289 講演の約 2.9M 語から学習した。これらの学習セットにはテストセットと共通の話者は含まれていない。認識タスクは 3 単語分の音声であるが、音響特徴量の抽出の際、その 3 単語を含む発話毎に CMS を行った。デコーダには HTK を用いた。言語重みは予備実験より 10 とした。

2.3　被験者による認識実験

実験は候補単語の検索機能を備えた GUI を通して行った。被験者は同じ音声を何度でも繰り返し聞くことが出来るが、一度単語を決定し次の課題に進むと、前の課題には戻れないこととした。音声サンプルはランダムに選択されているため、前後 1 単語より長いコンテキストを用いた判断は不可能である。被験者は、与えられた選択肢中に正解単語が存在しない場合もあることを知っており、また適切な単語が選択肢中に見つからない場合でも、一番近いと思う単語を選択するように指示されている。

被験者は 15 人からなり、3 人ずつ 5 グループに分けた。内訳は男性が 14 名女性が 1 名で、本学の学部 4 年以上の学生および教職員である。評価用サンプルを 5 分割し、各グループ内の 3 人に、同一の 100 単語を認識タスクとして与えた。1 つのサンプルに

∗ Assessment of the performance of automatic spontaneous speech recognition system using human recognition performance.

By Takahiro Shinozaki and Sadaoki Furui, Department of Computer Science Tokyo Institute of Technology

対し 3 人の被験者を用いたのは、人による単語の選択結果には、単純な選択ミスや各単語に対する親密度の個人差の影響が含まれると考えられるためである。これらの影響を取り除いた認識率の上限を推定し、人による認識率の上限値をデコーダに対する基準として用いることとする。このために、3 人の被験者による認識結果の単語間で多数決を行った結果を、最終的な人による認識結果として用いた。3 人の被験者の結果が全て異なる場合は、3 人の中で認識率が一番高い人の結果を用いた。被験者は 10 サンプルからなる練習用タスクを行った後、評価用タスクの認識を行っている。

3 実験結果

3.1 デコーダと人の比較

デコーダ (ASR) および人 (HSR) による認識率をグループ毎に図 2 に示す。認識率は未知語を除いた課題中で正しく認識された単語数の割合とした。デコーダと人で共通して、1 サンプルに対して必ず 1 単語選択することとしたため、挿入誤りや削除誤りは存在しない。デコーダによる認識では、音響モデルに不特定話者モデル (SI) を用いた場合に加え、参考として、それぞれの話者の講演全体を用いて、MLLR により教師無し話者適応 [1] を行ったモデル (SA) を用いた結果も合わせて示した。人による認識率は 95.3%、不特定話者音響モデルを用いたデコーダの認識率は 88.7%であり、母比率の差の検定を行うと 1%水準で有意であった。音響モデルに教師無し話者適応モデルを用いた場合のデコーダの認識率は 91.3%であり、人による認識率との差は 5%水準で有意であった。

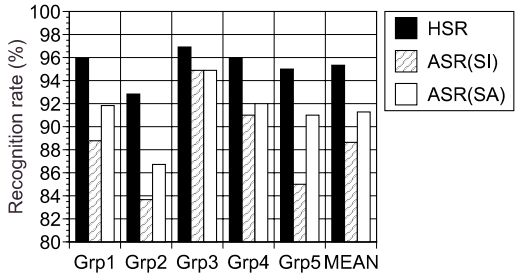

図 2. Human and decoder recognition rate.

デコーダと人の認識結果の共通部分および相違部分を表 2 に示す。デコーダが正しく人が間違った部分には、曖昧な発音で判別が難しいがデコーダの結果は合っていた場合や、2 人以上の被験者が同じ選択ミスをした場合が含まれる。

表 2. デコーダと人の正誤によるサンプルの分類

		ASR	
		Correct	False
HSR	Correct	426	45
	False	12	11

未知語: 6

表 2 においてデコーダが間違えて人が正しかった部分の 45 サンプルのうちの 33 は、3 人の被験者とも正解したサンプルであった。この部分を正しく認識できるようにすることで、6%程度の認識率の改善が見込める。これらの単語についてデコーダが認識に失敗した理由を調べるため、正解単語と認識結果単語の言語尤度、音響尤度の比較を行った。比較は、言語尤度は認識時の言語重みを含め、また音響尤度は前後単語の尤度も含めて行った。

正解単語と比べて音響尤度は低いが言語尤度が高かった 13 単語を詳しく見ると、正解単語列よりも出現しそうにない単語連鎖に高い尤度が割り当てられているものと、正解単語列の方がたまたま稀な単語連鎖である場合がほぼ半々であった。前者の方は、確率値の計算の際にバックオフが起きた部分で不自然な尤度が割り当てられていることから、学習データの不足が原因と考えられる。後者に関しては、正解単語の言語尤度が低いこと自体は問題ないことから、それを補うような音響尤度の改良も必要と考えられる。逆に、正解単語と比べて言語尤度は低いが、音響尤度が高かった 9 単語を見てみると、単語全体が無声化しているもの、小さなノイズが重畳しているものが 1 つずつあり、残りの 7 単語に関しては、音声に特に問題は見られないものと、3 単語を続けて聞くと特に問題なく聞こえるものの、中心の 1 単語のみ抜きだして聞こうとすると正解単語とやや異って聞こえるものがそれぞれ半数ほど含まれていた。

3.2 連続音声認識との関係

前後の単語コンテキストが与えられた条件での単語認識と連続音声認識を行った場合のデコーダの認識率の関係を調べるため、音響モデルや言語モデルの精度を様々にかえて実験を行った。結果を図 3 に示す。実験は表 1 の講演中の 3000 単語および 280 文を用いて行った。連続音声認識の結果は単語正解精度を用いて評価した。なお、認識結果の集計において表記の揺れを考慮していないことや、読み付与誤りがある部分を除外してないことなどにより、図 2 よりも単語認識率は低くなっている。単語認識率と連続音声認識率の関係は、ほぼ傾きが 1.72 で、 $(100\%, 96\%)$ となる点を通る直線で表せることが分かる。仮に単語認識において 6%認識率を向上させることが出来れば、連続音声認識の認識率は 10%向上すると予想される。

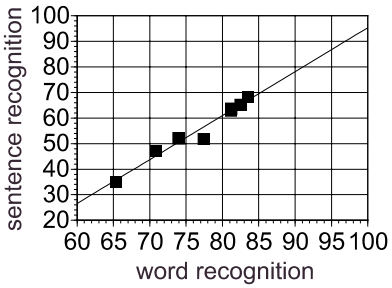

図 3. Word vs. sentence recognition rate.

4 まとめ

人の認識性能を基準としてデコーダの認識性能の評価を行った。人にとって問題なく認識できるにも関わらず、デコーダでは認識できない単語が約 6%存在すること、その原因として、言語モデルや音響モデルの推定精度、発音の変形や曖昧化が挙げられることを示した。この部分を改良できれば、連続音声認識で 10%程度の認識率の向上が期待される。

参 考 文 献

[1] 篠崎 隆宏, 古井 貞熙, “話し言葉音声認識における話者間の認識率変動要因の解析,” 信学技報, SP2001-102, pp. 1-6, (2001-12).