第9回：音声認識システムの評価指標

WER, CER, SLUE ベンチマーク

小野原彩香｜北星学園大学

この回は動画回です。Google Forms でテストを受けてください。

本日の到達目標

1. 音声認識システムの全体構成を整理し、各要素の役割を説明できる
2. WER と CER を数式に基づいて**正確に計算**できる
3. 編集距離（Levenshtein 距離）と WER/CER の関係を理解する
4. SLUE ベンチマークの概念を理解し、評価指標の使い分けを説明できる

音声認識システムの全体像

構成	音声から文字列への流れ
HMM系	特徴量 → 音響モデル。デコーダが発音辞書・言語モデルも組み合わせて語列を探索
End-to-End系	波形・特徴量 → エンコーダ → CTC、Transducer、AEDなどで文字・サブワード列を推定
評価	WER / CER、用途に応じたSLUEなどの理解タスク

CTC・AEDはモデル／学習方式、Viterbi・ビーム探索は探索アルゴリズム。

wav2vec 2.0・HuBERTには認識用の出力層などを追加して微調整する。外部言語モデルは任意に併用できる。

モデルの構成と学習方法

設計・学習方法	モデル例	特徴
HMM系	GMM-HMM / DNN-HMM	音響・発音辞書・言語モデルを統合
End-to-End	CTC / Transducer / AED	手作業のフレーム対応ラベルは不要
自己教師あり事前学習	wav2vec 2.0 / HuBERT	音声表現を学び、認識用に微調整
大量の音声・テキスト対	Whisper	多言語・多様な入力条件で学習

WERの比較には、評価分割・学習データ・言語モデル・正規化・探索条件をそろえる。

言語モデルとデコーダとの統合

$$\mathrm{score}(Y;X) = \log P_{\mathrm{ASR}}(Y \mid X) + \lambda \log P_{\mathrm{LM}}(Y)$$

λ：外部LMの重み。上式は探索スコアであり、正規化された対数確率ではない。

ビーム探索（Beam Search）

最もスコアの高いテキスト列を効率的に探索するアルゴリズム。

貪欲探索：各出力ステップで最高確率の1候補を選ぶ → 系列全体の最適解とは限らない

ビーム探索（ビーム幅 B）：上位 B 個の候補を常に保持しながら展開

```
ビーム幅 B=3 の例：

t=1: "a"(0.7)、"b"(0.2)、"c"(0.1) を保持

t=2: 各候補を展開 → 9候補 → 上位3を残す
   "ab"(0.42)、"ac"(0.21)、"ba"(0.10)、...

最終: 最もスコアが高い系列を出力
```

> B が大きいほど良い解を見つけやすいが、計算コストも増大する。

WER（Word Error Rate）とは

音声認識の最も基本的な評価指標——**単語レベルの誤り率**。

$$\boxed{\mathrm{WER} = \frac{S + D + I}{N} \times 100\%}$$

記号	意味	例
S	Substitutions（置換）	正解単語が別の単語に認識された
D	Deletions（削除）	正解単語が認識されなかった
I	Insertions（挿入）	存在しない単語が認識に現れた
N	正解文の総単語数	—

> 注意：$\mathrm{WER} > 100\%$ になることもある（挿入が多い場合）。

WER 手計算例①

正解文： 今日 は 天気 が いい （5単語）
認識結果： 今日 は 天候 いい （4単語）

アライメント：

```
正解:   今日  は  天気  が   いい
認識:   今日  は  天候       いい

今日 → 今日  (一致)
は   → は    (一致)
天気 → 天候  (置換: S+1)
が   → (なし) (削除: D+1)
いい → いい  (一致)
```

$$S = 1, \quad D = 1, \quad I = 0, \quad N = 5$$

$$\mathrm{WER} = \frac{1 + 1 + 0}{5} = \frac{2}{5} = \mathbf{40\%}$$

WER 手計算例②（挿入あり）

正解文： 今日 は 晴れ （3単語）

認識結果： 今日 は 今日 は 晴れ （5単語）

```
正解:   今日   は   晴れ
認識:   今日   は   今日   は   晴れ

「今日 は」が余分に挿入された
S=0, D=0, I=2, N=3
```

$$\mathrm{WER} = \frac{0 + 0 + 2}{3} = \frac{2}{3} \approx \mathbf{66.7\%}$$

> 挿入が多いと WER は 100% を超えることもある。

編集距離（Levenshtein 距離）

WER/CER の計算基盤は**編集距離**（最小操作回数で2つの列を一致させる）。

$$\mathrm{WER} = \frac{\text{編集距離（単語レベル）}}{\text{正解の総単語数}}$$

動的計画法による計算（$O(MN)$）：

$$d[i][j] = \min \begin{cases} d[i-1][j] + 1 & \text{（削除：ref の } i \text{ 番目を削除）} \\ d[i][j-1] + 1 & \text{（挿入：hyp の } j \text{ 番目を挿入）} \\ d[i-1][j-1] + \mathbf{1}[r_i \neq h_j] & \text{（一致 or 置換）} \end{cases}$$

初期条件：$d[0][j] = j$、$d[i][0] = i$

編集距離 DP テーブルの例

ref = `cat` **、hyp =** `car`

		c	a	r
	0	1	2	3
c	1	**0**	1	2
a	2	1	**0**	1
t	3	2	1	**1**

編集距離 = **1**（`t → r` の置換1回で一致）

→ 文字レベルの CER を計算する場合も全く同じ手順

CER（Character Error Rate）

文字レベルの誤り率——日本語などでよく使われる。

$$\boxed{\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c} \times 100\%}$$

指標	対象言語	使う理由
WER	英語・スペース分かち書き言語	分かち書き規則を固定して単語の誤りを測る
CER	日本語・中国語など	単語分割の規則の違いを避けて評価できる

> 日本語の「今日は良い天気だ」を単語に分けると形態素解析が必要。
> CERなら単語分割を避けられる。句読点・空白・表記の正規化はどちらも統一する。

CER 手計算例（テスト出題範囲）

正解文：今日は天気がいい（8文字）
認識結果：今日はてんきがいい（9文字）

文字単位のアライメント：

```
正解:   今 日 は 天 気 が い い
認識:   今 日 は て ん き が い い

操作:
「天」→「て」   置換  (S+1)
「気」→「ん」   置換  (S+1)
（なし）→「き」 挿入  (I+1)
「が」以降は一致
```

$$S_c = 2, \quad D_c = 0, \quad I_c = 1, \quad N_c = 8$$

$$\mathrm{CER} = \frac{2 + 0 + 1}{8} = \frac{3}{8} = \mathbf{37.5\%}$$

SLUE ベンチマーク

SLUE（Spoken Language Understanding Evaluation）
Shonほか（2022）が提案した、英語自然発話を用いる評価ベンチマーク。

評価	測るもの	主な指標
ASR	書き起こしの正確さ	WER
固有表現認識（NER）	人名・地名・組織名などの抽出	F_1
センチメント分析（SA）	ポジティブ・ネガティブ・ニュートラルの分類	**macro F_1**など

初期版は**2つのSLUタスク＋ASR**。Phase-2では課題が拡張されている。

> 書き起こしの正確さと下流タスク性能は別に評価する。SAの分類成績から、感情や意味理解全体の正確さは結論できない。

出典：SLUE原論文・公式評価方法

混同行列による誤り分析

音素混同行列：どの音素が他の音素に誤認識されやすいかを可視化。

```
音素混同行列の例（行=正解音素、列=認識音素）
          /a/   /i/   /u/   /e/   /o/
正解/a/ [  95    2     1     1     1   ...]
正解/i/ [   3   90     3     2     2   ...]
正解/u/ [   2    4    88     2     4   ...]
（架空の例。行ごとに正規化した割合 %）
対角成分が高いほど一致が多い。挿入・削除も別に数える
```

日本語でよくある誤認識例

正解単語	よくある誤認識	原因
橋（はし）	箸（はし）	音素列は同じ。アクセントや文脈の手がかりも条件に依存
音楽	音学	同じ読みからの表記選択の誤り（原因は要検証）

エラー分析の方法論

Step 1：量的分析

- WER/CER の全体値を計算
- 条件別（騒音・話者・発話速度）に分類

Step 2：エラーの分類

エラー種別	対処法
音響的誤り（類似音の混同）	音響モデルの改善・データ増強
語彙外単語（OOV）	辞書拡張・文字モデル使用
言語モデルエラー	LM 更新・ドメイン適応
ハルシネーション	無音・分割・復号条件を確認し、対策後も評価

Step 3：デバッグ → Step 4：改善策の立案と評価

Python 実装：WER/CER の計算

```
!pip install "jiwer>=4,<5"
from jiwer import wer, cer
ref_en = "the cat sat on the mat"
hyp_en = "the cat sat on the mat at"
print(f"WER: {wer(ref_en, hyp_en):.1%}")  # 16.7%（挿入 1 / 参照 6 単語）
ref_ja = "今日は天気がいい"
hyp_ja = "今日はてんきがいい"
print(f"CER: {cer(ref_ja, hyp_ja):.1%}")  # 37.5%（編集 3 / 参照 8 文字）
```

自前の実装（最小編集距離を動的計画法で求める）

```
def edit_distance(reference, hypothesis):
    m, n = len(reference), len(hypothesis)
    dp = [list(range(n + 1))] + [[i] + [0] * n for i in range(1, m + 1)]
    for i in range(1, m + 1):
        for j in range(1, n + 1):
            cost = int(reference[i - 1] != hypothesis[j - 1])
            dp[i][j] = min(dp[i - 1][j] + 1,       # 削除
                           dp[i][j - 1] + 1,       # 挿入
                           dp[i - 1][j - 1] + cost) # 一致 / 置換
    return dp[m][n]
```

本日のまとめ

1. **WER** = $\frac{S+D+I}{N}$ ：単語レベルの誤り率。英語の標準指標。
2. **CER** = $\frac{S_c+D_c+I_c}{N_c}$ ：文字レベルの誤り率。**日本語でよく用いられる指標**。
3. **編集距離**：WER/CER の計算基盤。動的計画法で $O(MN)$。
4. **SLUE ベンチマーク**：NER・感情分析による意味理解の評価。WER/CER を超えた評価。
5. **エラー分析**：量的分析 → 分類 → デバッグ → 改善策立案。

第7〜9回の流れ

```
第7回: 音声と統計モデル（GMM-HMM, wav2vec 2.0）
  ↓ モデルを学んだ
第8回: 最新 E2E モデル（Conformer, Whisper）と実習
  ↓ 使い方を学んだ
第9回: 評価指標とエラー分析
  ↓ 評価方法を学んだ
第10回: 音声インタフェースと UX 設計 →
```

Google Forms テストのご案内

第9回テスト（5問・選択式）

テストは別途メールでお知らせしたURLから受験してください。

テスト問題 1：WER の計算

正解文：私 は 毎日 学校 へ 行く （6単語）
認識結果：私 は 毎日 学校 行く （5単語）

このときの WER はいくらか？

1. 1/6 ≈ 16.7%
2. 2/6 ≈ 33.3%
3. 1/5 = 20.0%
4. 0%

テスト問題 1：解説

```
正解:   私  は  毎日  学校  へ  行く
認識:   私  は  毎日  学校      行く

「へ」が削除されている → D=1, S=0, I=0
N=6（正解の総単語数）
```

$$\mathrm{WER} = \frac{0 + 1 + 0}{6} = \frac{1}{6} \approx 16.7\%$$

正解：選択肢 1

> 選択肢 3（1/5）は誤り——分母は**正解文**の単語数（6）であり、認識結果の単語数（5）ではない。

テスト問題 2：CER の計算

正解文：今日は天気がいい（8文字）
認識結果：今日はてんきがいい（9文字）

このときの CER はいくらか？

1. 1/8 = 12.5%
2. 2/8 = 25.0%
3. **3/8 = 37.5%**
4. 4/8 = 50.0%

正解：選択肢 3（動的計画法で $S_c = 2, D_c = 0, I_c = 1, N_c = 8$）

テスト問題 3：WER と CER の使い分け

単語分割の規則の違いを避けたいとき、CERを選ぶ理由として**最も適切**なのはどれか。

1. 英語の認識結果を評価するとき
2. **日本語の認識結果を評価するとき（単語分かち書きが難しいため）**
3. 認識結果に挿入エラーが多いとき
4. 言語モデルを使わない End-to-End システムを評価するとき

正解：選択肢 2

> 日本語・中国語などでは通常は単語間に空白を置かない。CERは単語分割の違いを避けられる。日本語でも分割規則を固定してWERを使える。

テスト問題 4：編集距離の理解

`reference = [A, B, C, D]`（4単語）、`hypothesis = [A, C, D, E]`（4単語）
の最小編集距離（単語レベル）はいくらか？

1. 1
2. **2**
3. 3
4. 4

正解：選択肢 2（B を削除 D+1、E を挿入 I+1 → 合計2操作）

$$\text{WER} = \frac{0 + 1 + 1}{4} = 50\%$$

テスト問題 5：SLUE ベンチマーク

SLUE ベンチマークに関する説明として**最も適切なもの**はどれか。

1. WER と CER を統一的に評価するためのベンチマーク
2. **音声言語理解（意味理解）の能力を評価するベンチマーク（NER・感情分析を含む）**
3. 日本語特有の評価指標で漢字変換精度を評価する
4. 大規模言語モデル（LLM）の音声理解能力を評価するベンチマーク

正解：選択肢 2

> SLUE初期版には、NER・センチメント分析に加えて、同じデータでのASR評価も含まれる。