15th Sept. 2016, ASJ Annual Meeting Autumn@Toyama Univ.

尤度ベース言語識別における待ち時間短縮法

T. Okamoto, A. Hiroe and H. Kawai, ”Reducing latency for language identification based on large-vocabulary continuous speech recognition,” *Acoust. Sci. & Tech.*, (accepted, to appear).

○岡本拓磨，廣江厚夫，河井恒

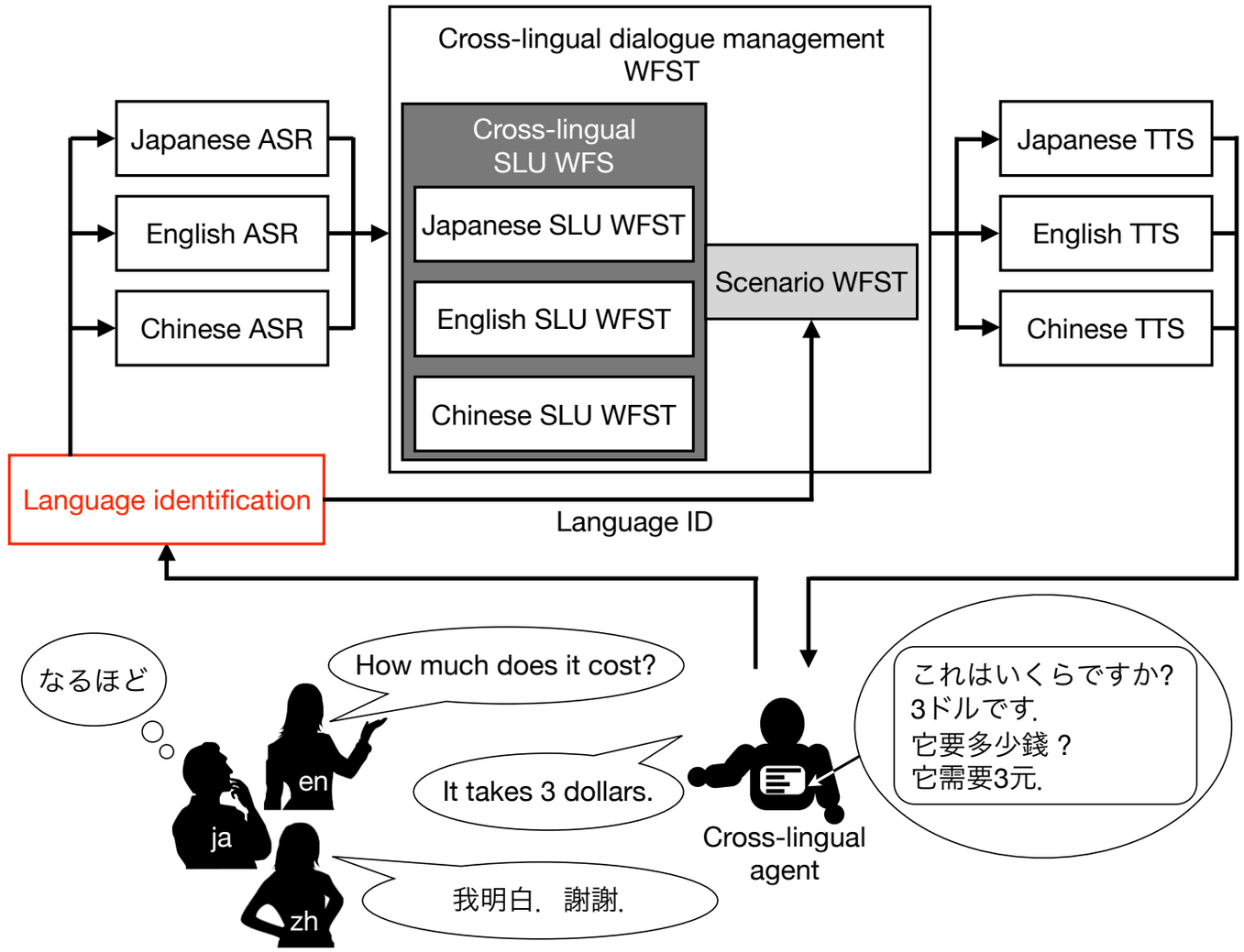
情報通信研究機構

Outline

- Introduction
- Comparison of previous methods
  - Discriminative approaches
  - Multiple large vocabulary continuous speech recognition approach
- Latency problem in LVCSR based LID
  - Latency problem
  - Timeout
  - Tradeoffs between accuracy and latency
- Proposed method
- Experiments
- Concluding remarks

Introduction

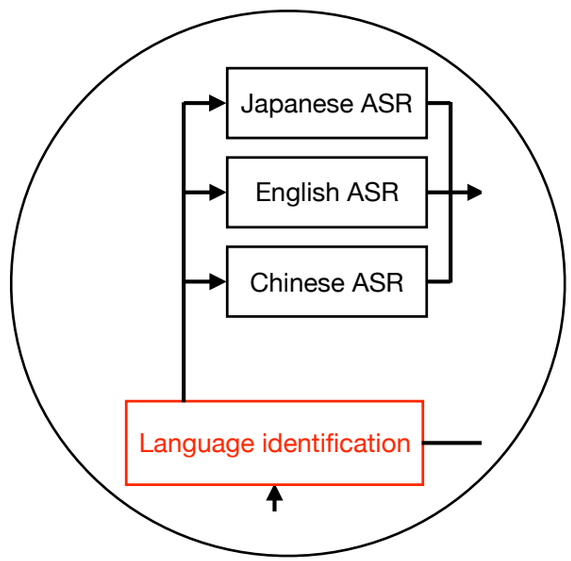
- 言語識別
  - 入力された言語が何語であるかを自動で推定する技術
    - Applications: VoiceTra，クロスリンガル音声対話システム..



岡本ら，音講論(秋)，2015

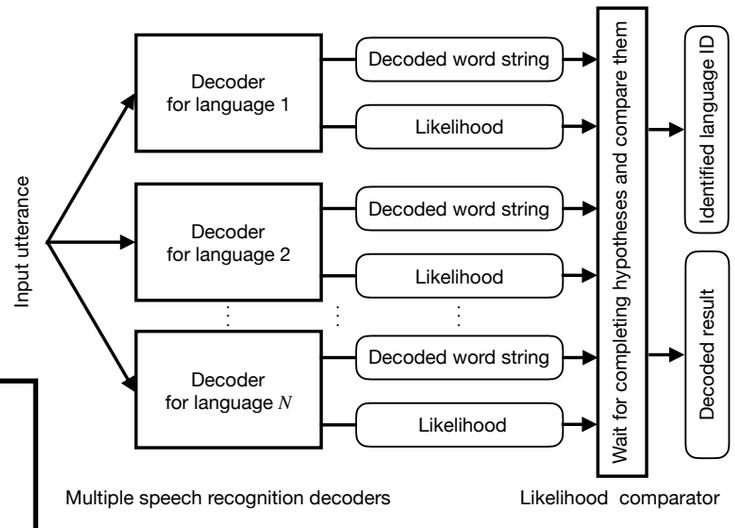
Comparison of previous methods

- 識別器による方式(現在の主流)
  - 入力音声特徴量から言語を直接識別 with i-vector and DNN
    - ○：20～30言語でも高精度に識別可能(研究ベース)
    - ×：識別学習要，言語追加には再学習要



- 尤度ベースによる方式(20年前に登場)
  - すべての音声認識器でデコードし，尤度最大を選択
    - ○(当時)：言語追加容易，実装が容易
    - ×(当時)：(20言語などの場合)精度は識別器に劣る
      計算コスト高

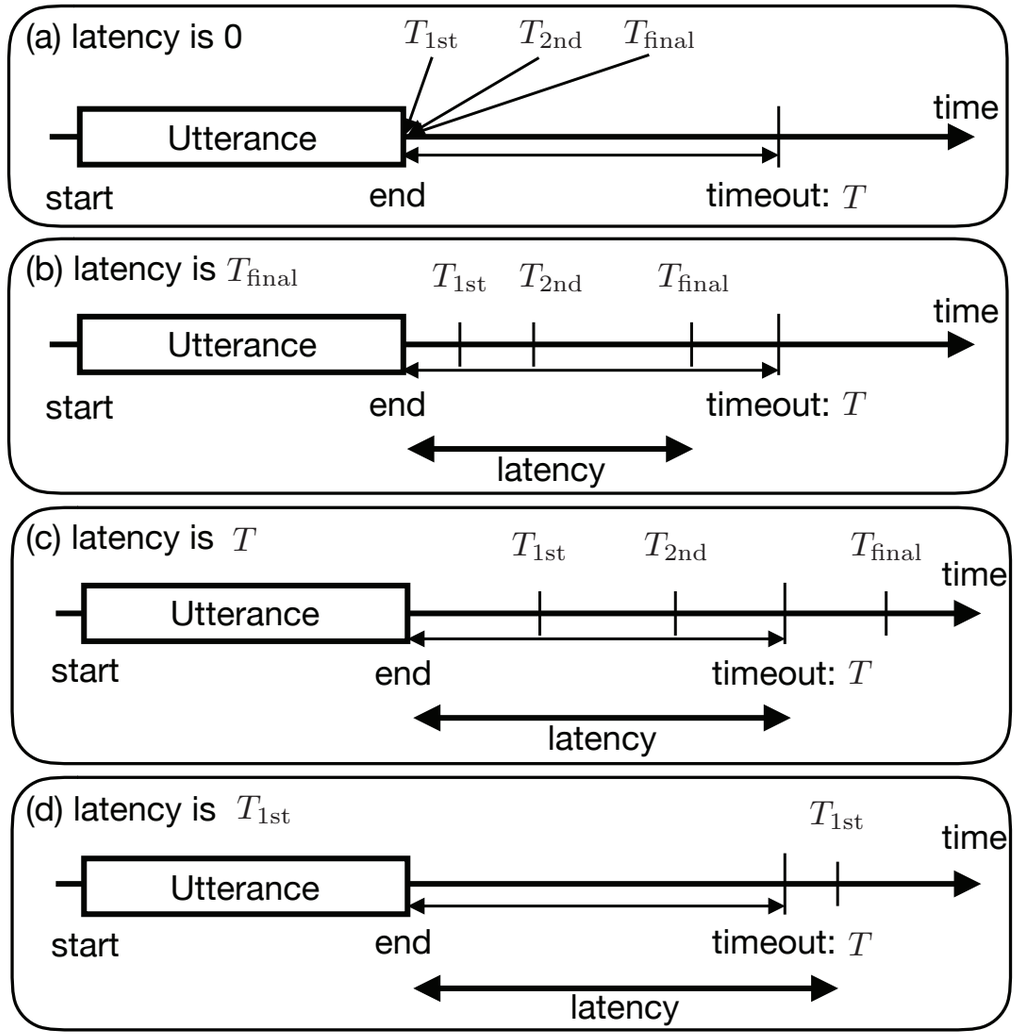
・深層学習による音声認識モデルの向上
・計算機性能の向上
・比較的少ない言語(～10)での識別



  - ○：識別学習不要，言語追加容易，実装&ラピッドプロトタイピング容易
  - ×：すべての認識結果を待つため，レイテンシーの発生

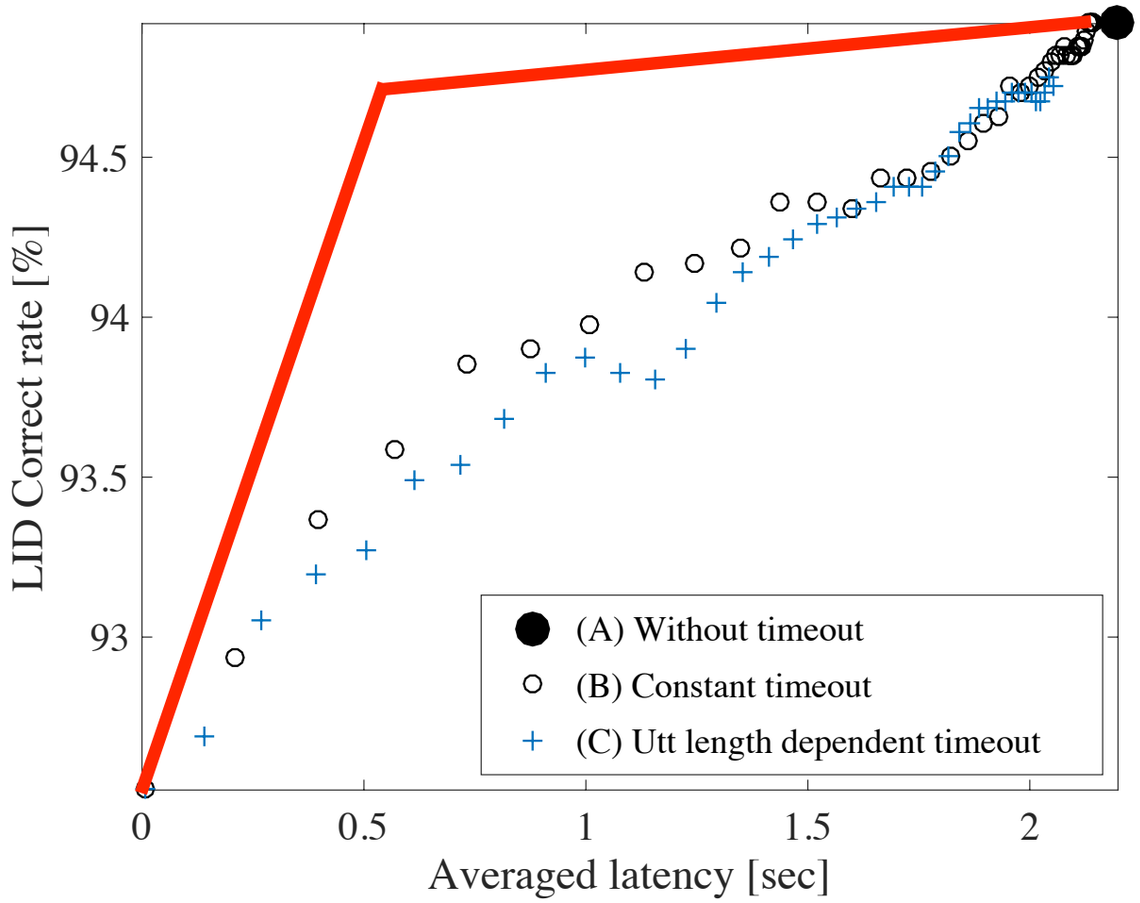
Latency problem in LVCSR based LID

- レイテンシー発生の理由
  - モデルミスマッチ
    - 入力とデコーダで言語が食い違う場合，モデルミスマッチのため仮説が膨大となり，言語が合っている場合と比べて計算時間がよけいにかかる
- タイムアウトの導入
  - 単純タイムアウト
    - 発話終了後，特定の時間内に届いた結果のみで尤度比較
  - 発話長に応じた可変型タイムアウト
    - 固定値ではなく，発話長に応じてタイムアウトを設定



Tradeoffs between accuracy and latency

- 識別精度とレイテンシーのトレードオフ問題
  - タイムアウトを小さくすると識別精度劣化
    =速いものを選択すればいいわけではない：何故か?



日英中3言語識別

目標：精度を保ちつつレイテンシーを小さくする方式の開発

Proposed method

- 速いものを選択すればよいわけでない理由
  - 各認識器ごとに処理速度が異なる
  - 速い認識器では言語が食い違っていても速く届く

	training data	vocablaly	averaged RTF
English (en)	370 h	262 k	0.47
Japanese (ja)	360 h	684 k	0.66
Chinese (zh)	560 h	213 k	0.27

入力発話言語

認識器言語	en	ja	zh
en	0.42	0.92	0.90
ja	1.55	0.67	1.44
zh	0.58	0.72	0.28

SPREDSの平均RTF結果

単純タイムアウトでは，ある識別器ではタイムアウト時間が長過ぎて，レイテンシーの原因とるが，別の識別器では短過ぎて正解がドロップアウト

- 各認識器の処理速度に応じた可変型タイムアウト
  - 各認識器の平均リアルタイムファクターRTF(発話長と処理時間の比)を使用し，認識器ごとにタイムアウト長を設定
    - ❋ 平均RTFはモデルリリースの際，容易に取得可能

$$T_{\mathrm{vt},l} = \begin{cases} w_{\mathrm{vt}}\mathrm{RTF}_l L - L & (\text{if } w_{\mathrm{vt}}\mathrm{RTF}_l > 1) \\ 0 & (\text{elsewhere}) \end{cases}$$

※$T_{\mathrm{vt,lang}} < 0$を防ぐため

Experiments

- 日英中3言語識別実験
  - DNN型音声認識(MCML音声インタラクションSDKモデル)
    - MFCC39次元(11フレーム=429次元)
    - 隠れ層5，512ノード
    - 音素数　英：39，日：26，中：30
  - 尤度を発話長で正規化

	training data	vocablaly	averaged RTF
English (en)	370 h	262 k	0.47
Japanese (ja)	360 h	684 k	0.66
Chinese (zh)	560 h	213 k	0.27

$$l = \underset{l \in \{\mathrm{en,ja,zh}\}}{\arg\min} \left(\frac{S_{\mathrm{utt},l}}{L_l} \right)$$

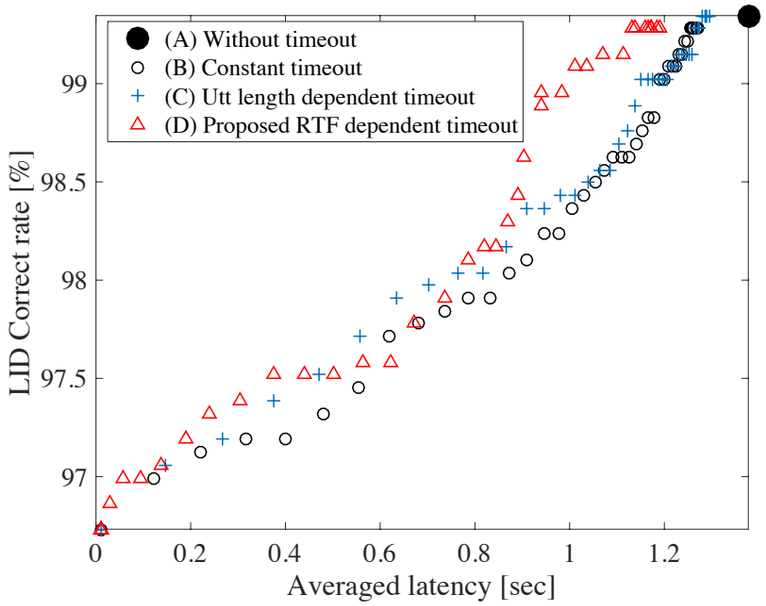
- テストセット

	en	ja	zh	L_{avg} (sec)	
BTEC	510	510	510	3.1	読み上げ音声，旅行会話，マイク収録
XTEC	1500	1500	912	4.6	模擬対話，マイク収録
VoiceTra	1726	5000	607	3.2	自然発話，スマホ収録，雑音込
SPREDS	1133	1498	1501	6.5	模擬対話，スマホ収録

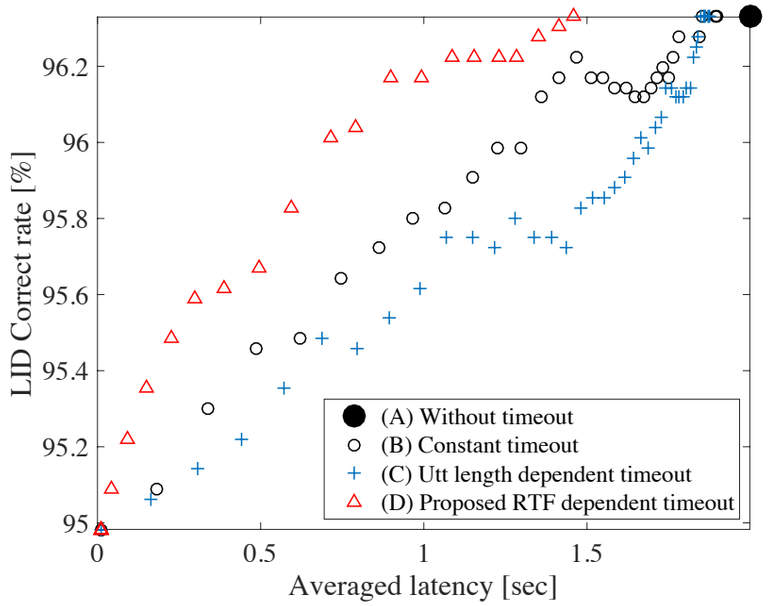
- 平均認識率　英：90.96 %　日：89.25 %　中：79.89 %

Results with BTEC, XTEC and VoiceTra

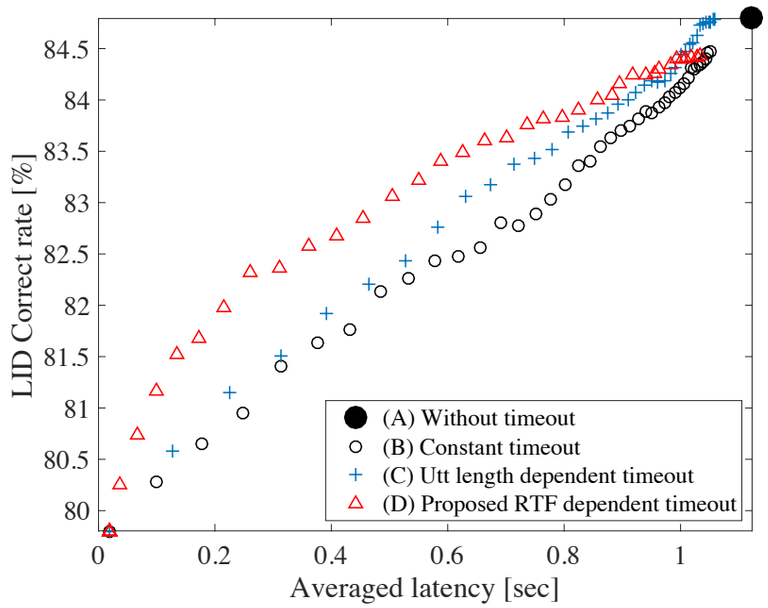
日英中3言語識別



BTEC



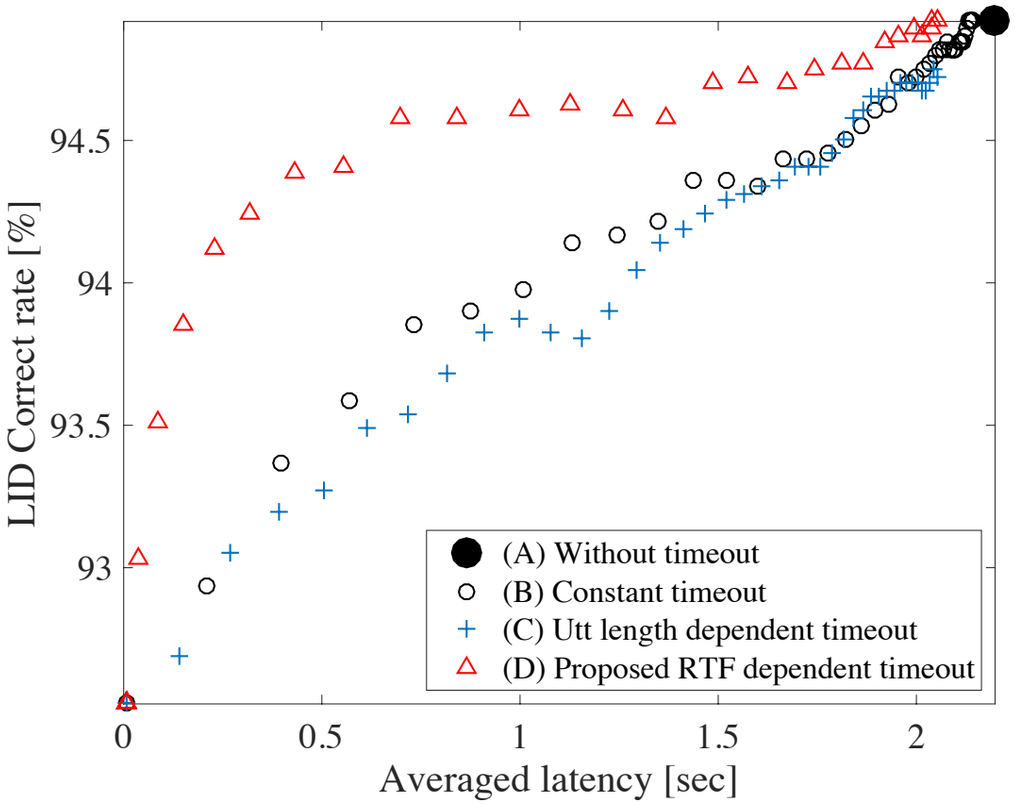
XTEC



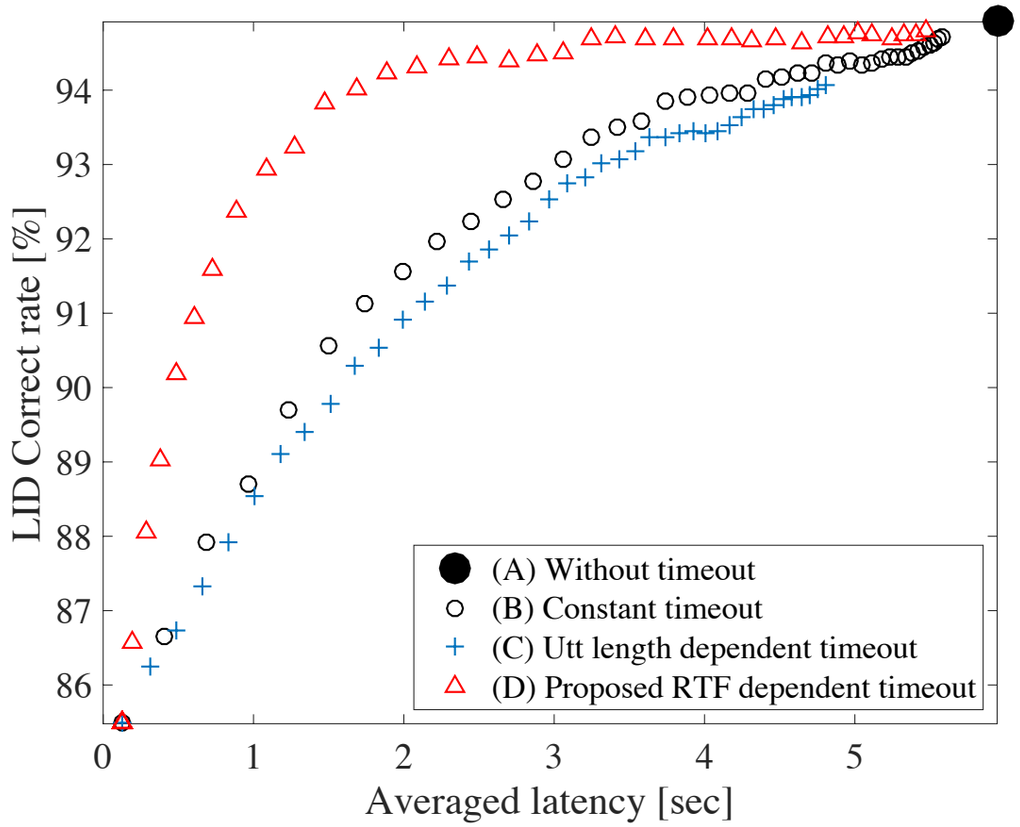
VoiceTra

Results with SPREDS corpus

日英中3言語識別



実測結果



1.5倍処理時間がかかった場合
(遅いマシンでの計算を再現)

精度を保ちつつレイテンシーを小さくする方式の実現

Concluding remarks

- 尤度ベース言語識別における待ち時間短縮法の提案
  - 尤度ベース言語識別の有用性
  - 待ち時間問題と単純タイムアウトのトレードオフ
  - 平均RTFを用いたデコーダごとの可変タイムアウト方式の提案
  - 日英中3言語識別実験による提案法の有効性の確認