1-2-5
統計的パラメトリック音声合成のための オーディオブックを用いた 学習コーパス自動構築

沢田慶，伊神和輝，浅井千明，佐藤雄介，
橋本佳，大浦圭一郎，南角吉彦，徳田恵一
名古屋工業大学

日本音響学会2016年春季研究発表 2016年3月9日

はじめに

- 統計的パラメトリック音声合成 (SPSS)
  - 音声データとテキストの対から音響モデルを学習
  - 大量の学習データを用いるほど音質が向上
- ビックデータを用いた統計モデル学習
  - 様々な研究分野で高い性能
  - SPSSのための学習データ
    - クリーンかつ同一環境で収録された音声が必要
      ⇒ 大量の音声データの収録には高コスト
- オーディオブックを用いたSPSSの学習
  - 比較的容易に利用できるビックデータ
  - 音声データの発話内容とテキストには不一致が存在

オーディオブックを用いた学習コーパスの自動構築

オーディオブック

- 書籍 (テキスト) を朗読したコンテンツ
  - 有料無料の大量のコンテンツがインターネット上に存在
  - 音声データとテキストの対が存在
  - 同一の話者による比較的クリーンかつ同一環境で収録

  ⇒ 音響モデル学習のためのビックデータとして有用

- オーディオブックを学習データとする問題点
  - 音声データの発話内容とテキストに不一致が存在
    - 置換: テキストの読み間違い
    - 削除: テキストに対応する音声の未収録
    - 挿入: テキストに対応しない音声の収録

音声データ	
ブックテキスト	he came to a cottage

⇒ 音響モデルの学習に悪影響

オーディオブックのテキスト

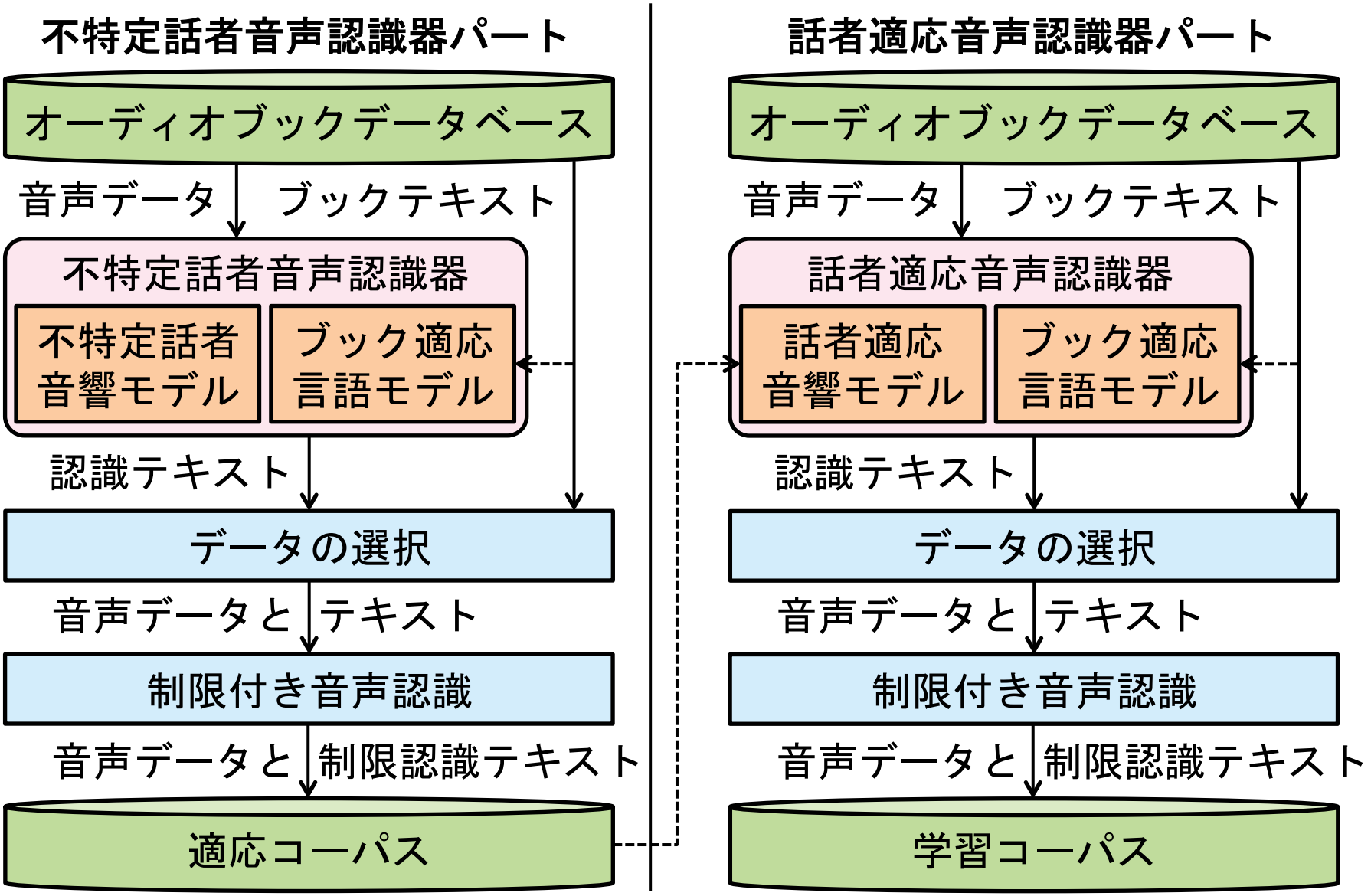
- 音声データの発話内容に一致する正解テキスト
  - 正解テキストを人手により作成することは高コスト
- 音声認識を用いた認識テキスト [Braunschweiler *et al.*; ’10]
  - 音声データの発話内容を推定
  - 認識エラーが含まれる可能性

音声データ	
正解テキスト	he came to a cottage knock knock
ブックテキスト	he came to a cottage _____
認識テキスト	she came to a cottage knock knock

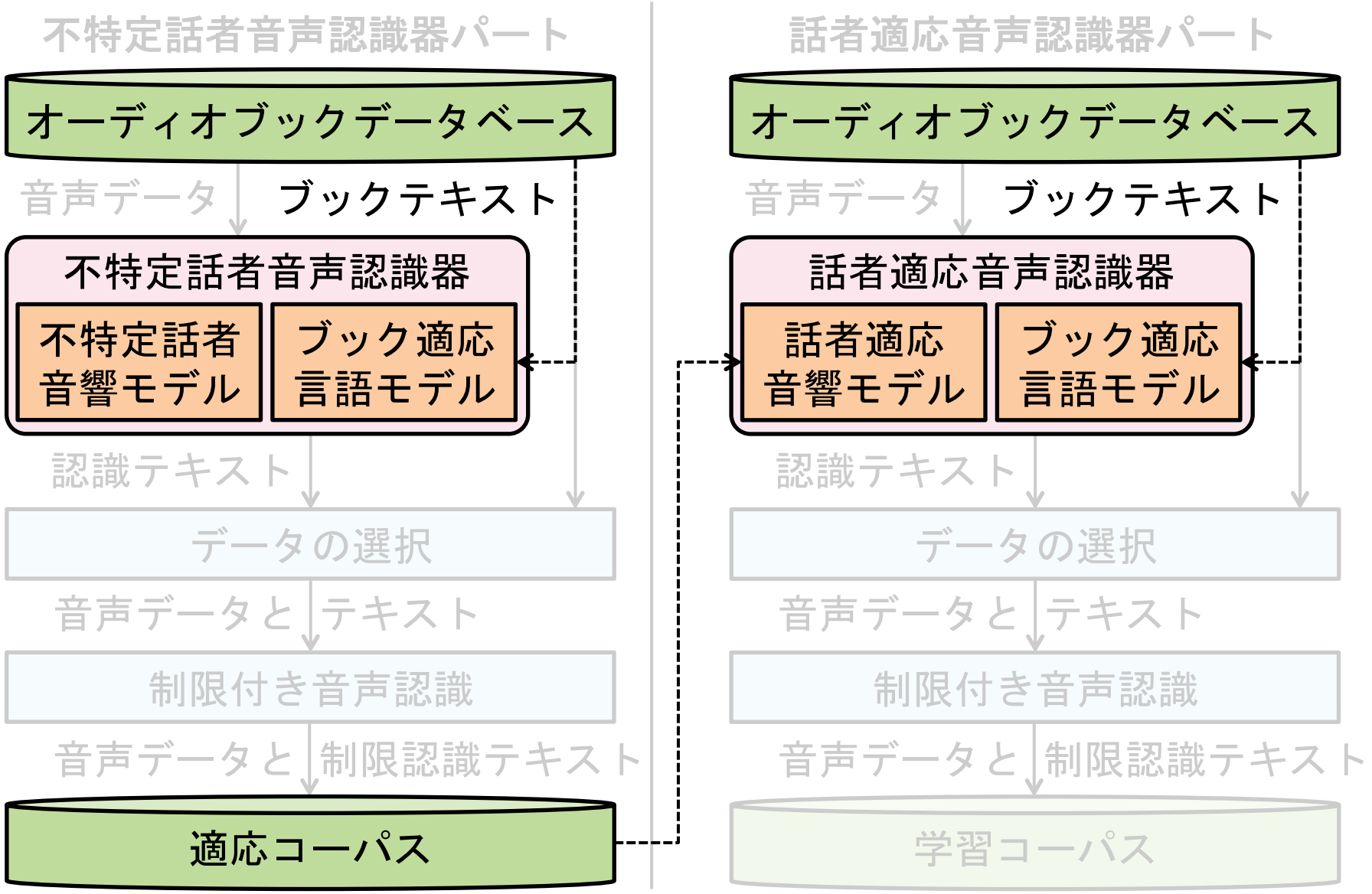
- 単語ネットワークに制限を設けた制限付き音声認識
  - ブックと認識テキストから単語ネットワークを構築
  - 認識エラーを削減し発話内容の不一致に対応可能

学習コーパスの質の向上を目指す

学習コーパス自動構築

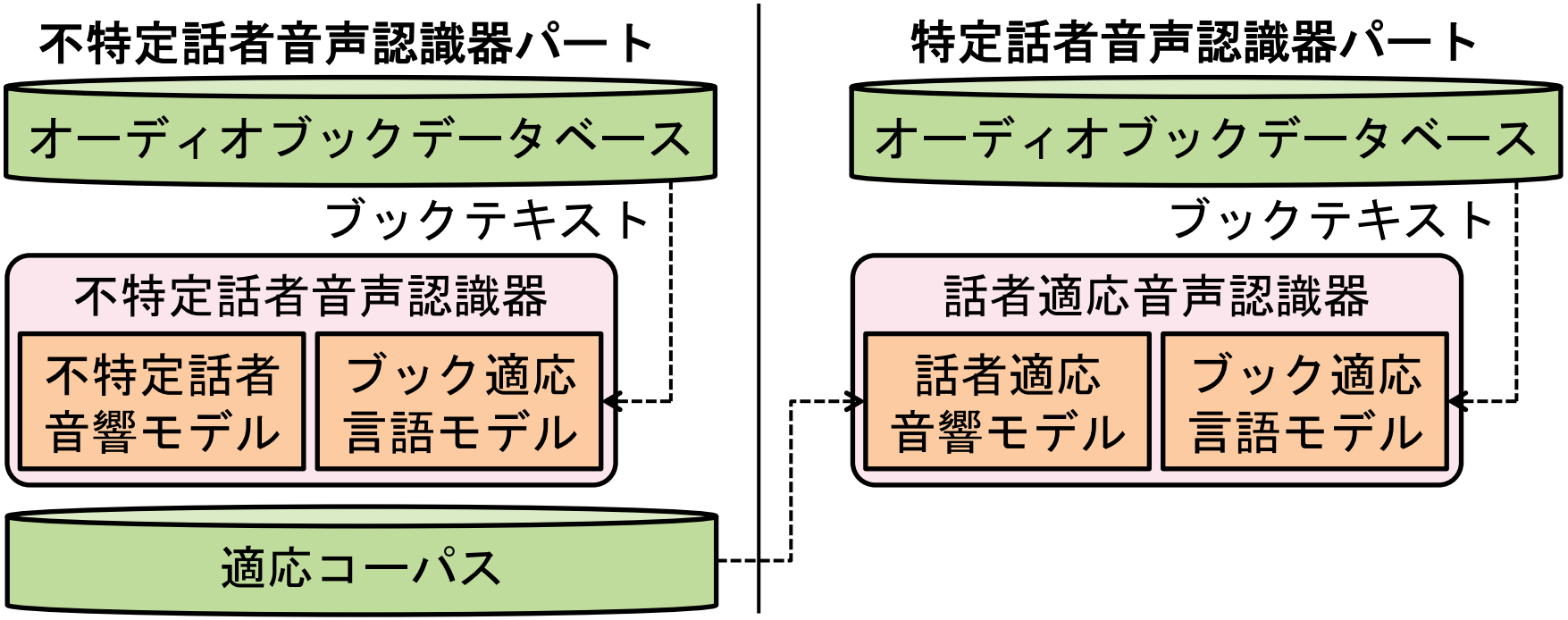

不特定話者音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
不特定話者音声認識器
不特定話者音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
適応コーパス
話者適応音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
話者適応音声認識器
話者適応音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
学習コーパス

学習コーパス自動構築

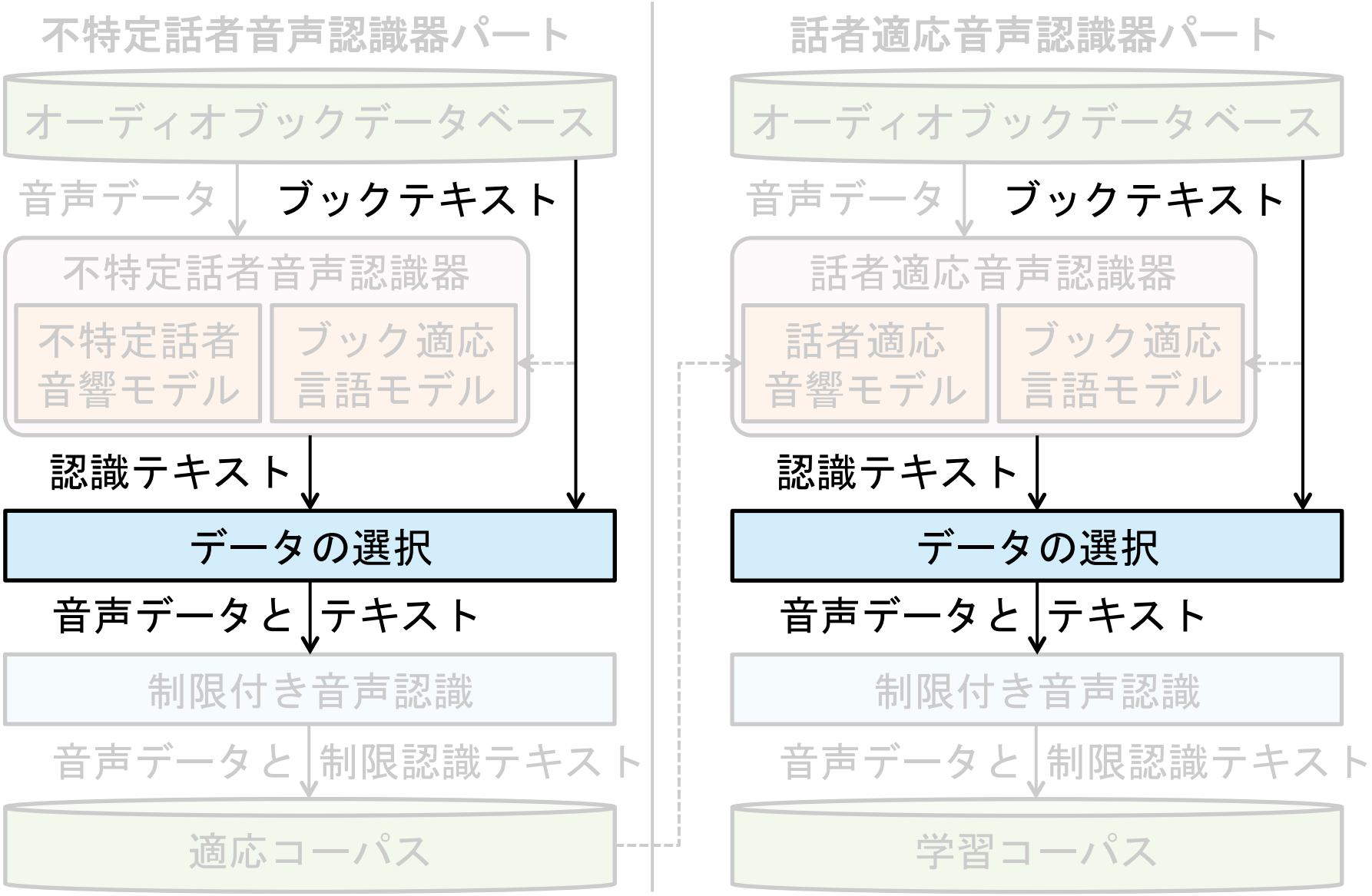

不特定話者音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
不特定話者音声認識器
不特定話者音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
適応コーパス
話者適応音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
話者適応音声認識器
話者適応音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
学習コーパス

オーディオブックを用いた音声認識の改善

- 言語モデル
  - 音声データの発話内容とブックテキストの大部分は一致
  - ブックテキストを用いた言語モデルは有用
  - ブックごとにブック適応言語モデルを構築
- 音響モデル
  - オーディオブックから話者適応音響モデルを構築



学習コーパス自動構築


不特定話者音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
不特定話者音声認識器
不特定話者音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
適応コーパス
話者適応音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
話者適応音声認識器
話者適応音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
学習コーパス

データの選択

- ブックテキストと認識テキストの比較

ブックテキスト	he　came to a cottage
認識テキスト	she came to a cottage knock knock

単語一致精度

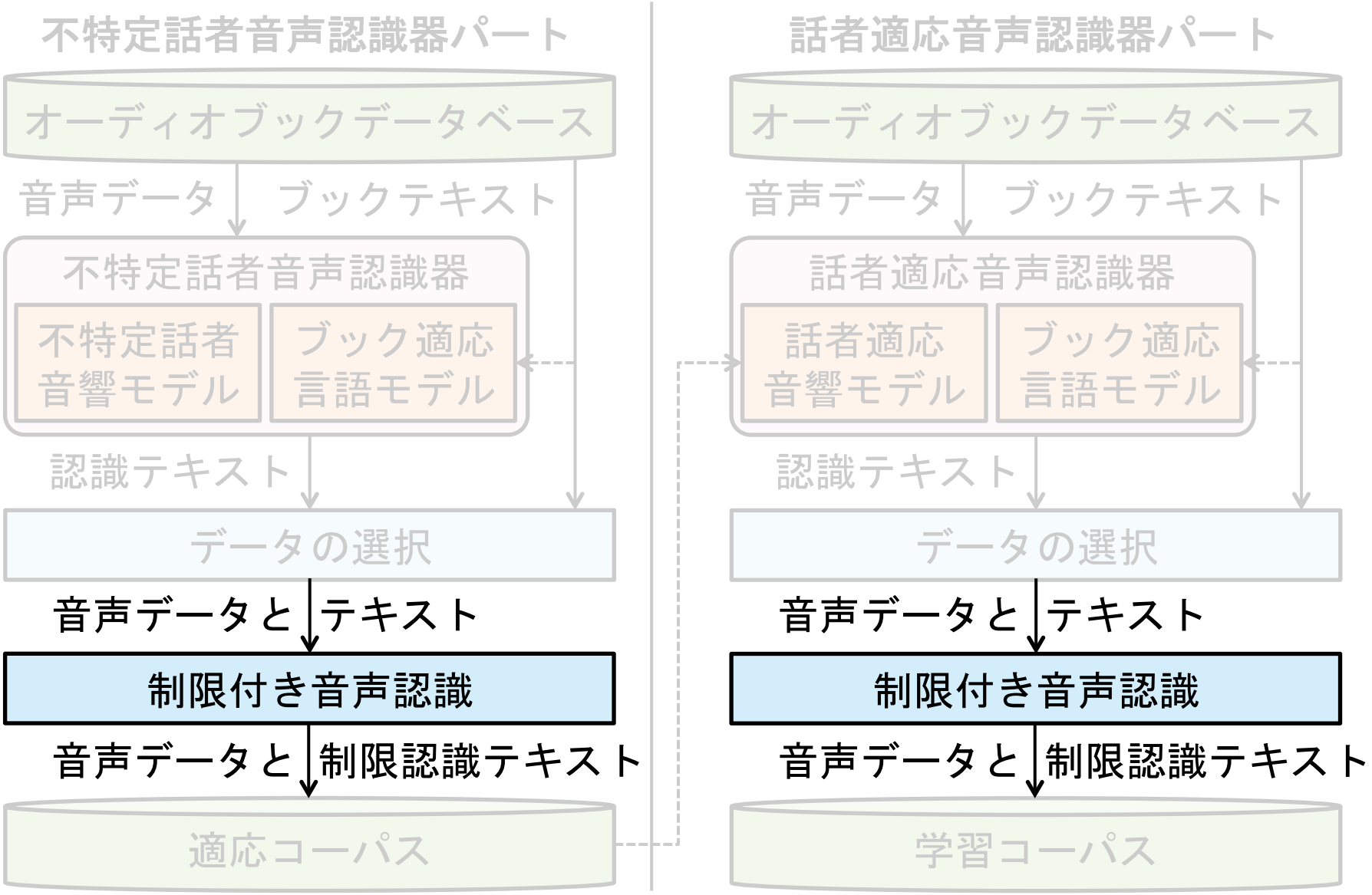
  - 単語一致精度が高い ⇒ 正解テキストと近い可能性大

- データの選択
  - 単語一致精度に閾値を設けデータを選択
  - 閾値の設定による学習コーパスの量と質はトレードオフ

閾値	大きい	小さい
学習コーパスの量	少ない	多い
学習コーパスの質	高い	低い

適切な閾値の設定により音響モデルの精度向上

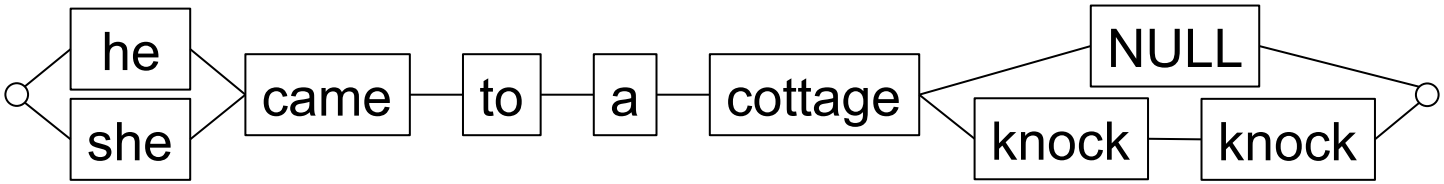
学習コーパス自動構築


不特定話者音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
不特定話者音声認識器
不特定話者音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
適応コーパス
話者適応音声認識器パート
オーディオブックデータベース
音声データ
ブックテキスト
話者適応音声認識器
話者適応音響モデル
ブック適応言語モデル
認識テキスト
データの選択
音声データと
テキスト
制限付き音声認識
音声データと
制限認識テキスト
学習コーパス

制限付き音声認識

- 制限付き音声認識による制限認識テキスト
  - ブックと認識テキストから単語ネットワークを構築
  - テキストが異なる場合は選択可能な単語ネットワーク
  - パスにペナルティを付加
    - ブックテキストがNULL: ブックテキストにペナルティ
      その他: 認識テキストにペナルティ
  - 言語モデルは利用しない音声認識

ブックテキスト	he　came to a cottage
認識テキスト	she came to a cottage knock knock



正解テキストに近いテキストを推定可能

実験条件 (1/3)

- オーディオブック

データベース	Usborne Publishing Ltd of audiobook recordings for Blizzard 2015
音声ファイル	ページ単位，平均16秒
学習データ	19冊，415ページ，115分
評価データ	3冊，48ページ

画像は http://www.usborne.com から引用

実験条件 (2/3)

- 音声認識器

不特定話者認識器構築データベース	TIMIT, WSJ0, WSJ1
サンプリング周波数	16 kHz
窓	ハミング窓
フレーム	長: 25 ms, シフト: 10 ms
特徴量	12次元MFCC + Δ + ΔΔ
音響モデル	3状態スキップなしleft-to-right HMM 16 (ポーズは32) 混合のGMM
言語モデル	bi-gram
適応コーパスとする単語一致度の閾値	90%
学習コーパスとする単語一致度の閾値	70%

実験条件 (2/3)

音声合成器

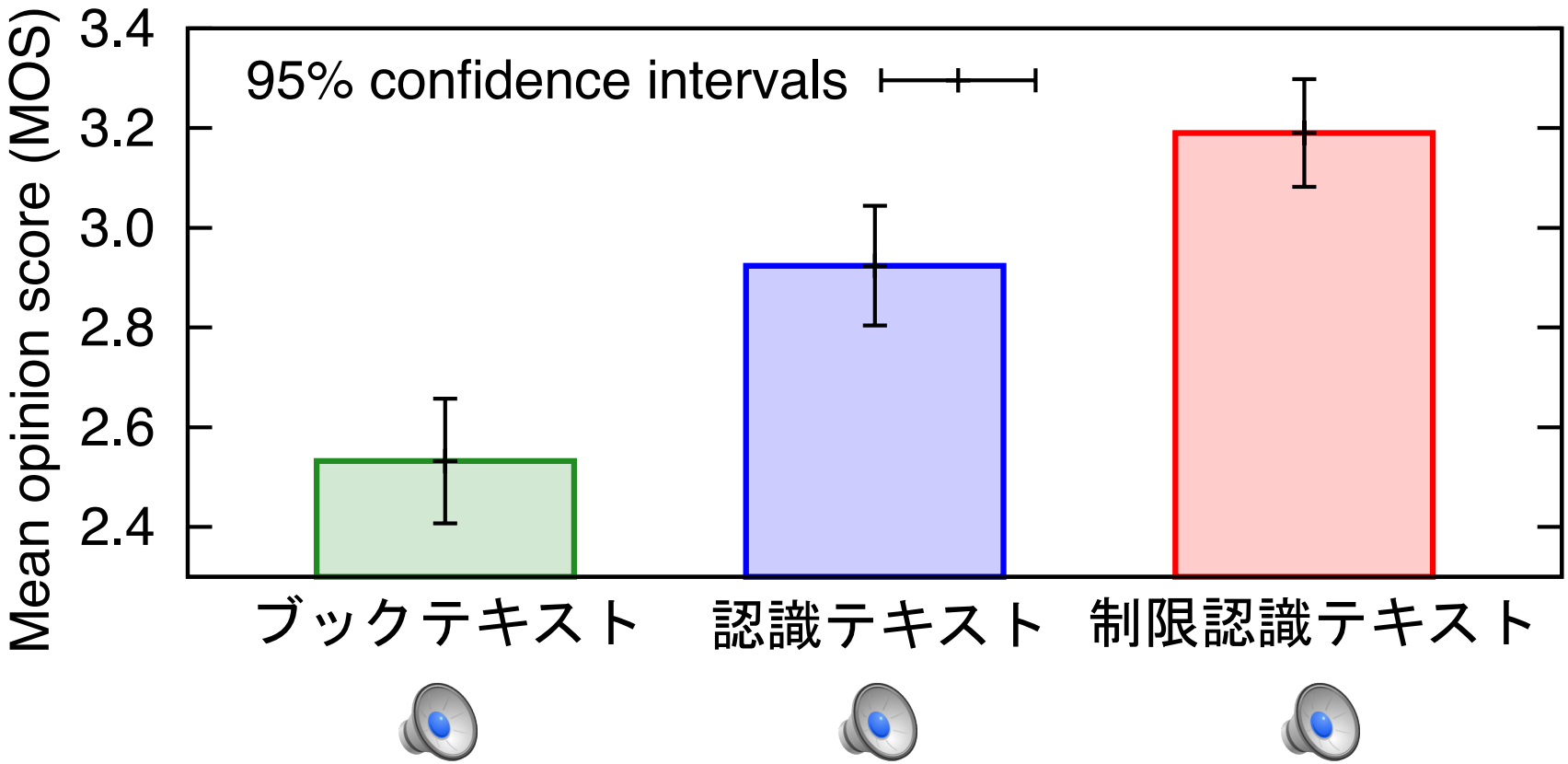
サンプリング周波数	44.1 kHz
窓	ガウシアン窓
フレーム	シフト: 5 ms
特徴量	39次元STRAIGHTメルケプストラム + Δ + ΔΔ, 19次元非周期成分 + Δ + ΔΔ, 対数基本周波数 + Δ + ΔΔ,
モデル構造	5状態スキップなしMSD-HSMM

評価実験

評価手法	自然性に関する5段階MOS試験
被験者数	8人
評価文章	評価データ3冊の中からランダムに20ページ

実験結果

- テキストの違いの比較



制限認識テキストは高い自然性

むすび

- オーディオブックを用いたSPSSの学習
  - 音声認識を用いた学習コーパスの自動構築
    - オーディオブックを用いた言語・音響モデル
    - データとテキストの選択
  - 主観評価実験
    - 単語ネットワークに制限を設けた音声認識は有効
- 今後の課題
  - 母語話者による主観評価実験
  - 大規模データベースを用いた実験
  - 表現豊かな音声合成システムの構築
- Blizzard Challenge 2016に向けて
  - ダブルクォート・構文解析の情報をコンテキストに導入
  - DNN音響モデルの導入 [Hashimoto *et al.*; ’16]