3-1-1
発音情報が未知の言語における テキスト音声合成システム構築法の検討

☆沢田慶，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2015年秋季研究発表 2015年9月18日

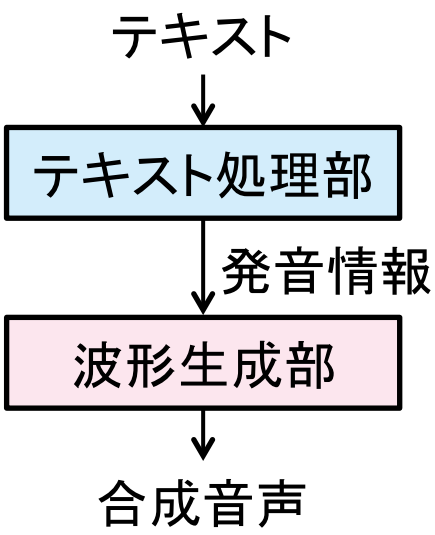
はじめに

- テキスト音声合成(TTS)システム
  - 任意のテキストの音声を合成するシステム
  - 様々なアプリケーションで利用
  - 高音質，多言語，発話スタイル等の需要が増加
- TTSシステムの多言語対応
  - 世界には数千言語が存在
  - あらゆる言語のTTSシステムを構築できる手法の確立
    ⇒ 音声合成研究の一つのゴール
  - ターゲット言語のTTSを構築には専門的な知識が必要
- Blizzard Challenge 2013-15 [Black, et al.; ’05]
  - インド語10言語のTTSシステムを構築するタスク
  - 音声データとテキストのみが提供

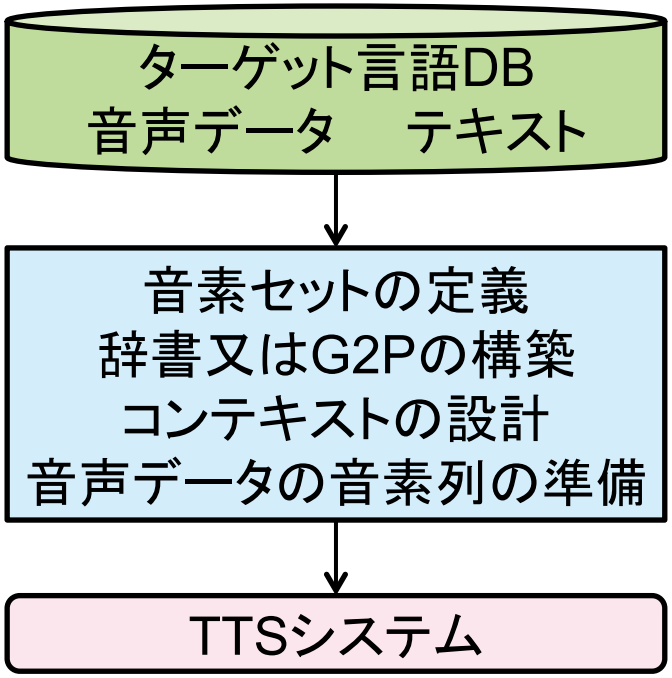
あらゆる言語のTTSシステムを構築すること目指す

TTSシステム

- TTSシステムの構成要素
  - テキスト処理部: 発音情報を推定
    - 発音情報が書かれた辞書を利用
  - 波形生成部: 発音情報に基づき波形を生成
    - 統計的パラメトリック音声合成等



- TTSシステムの構築



ターゲット言語に関する専門的な知識を利用した人手による作業
⇒ 発音情報が未知の言語のTTSを構築することは困難

専門的な知識を利用せずにTTSを自動構築する手法の検討

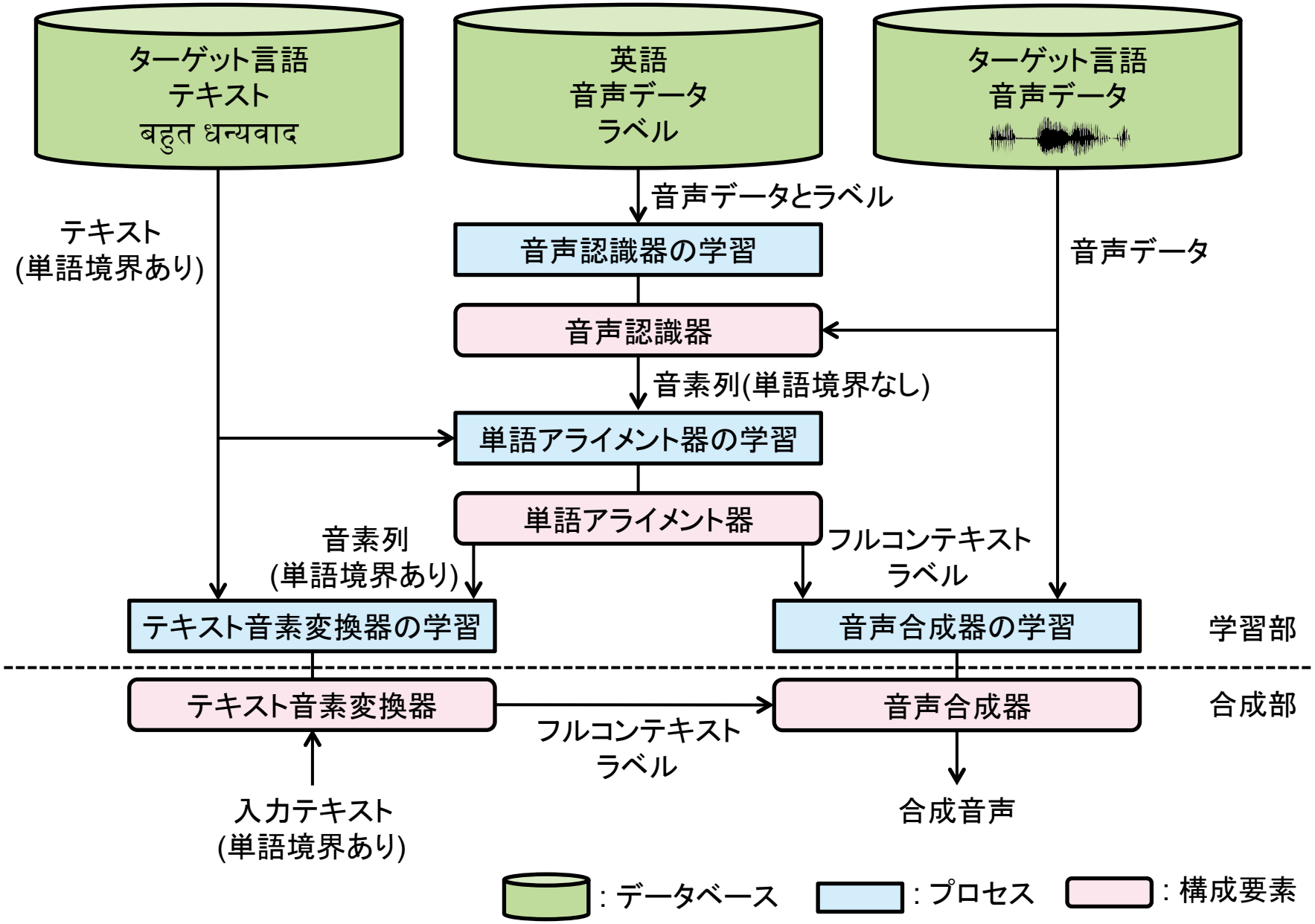
発音情報未知言語におけるTTSシステム構築

音声データ:

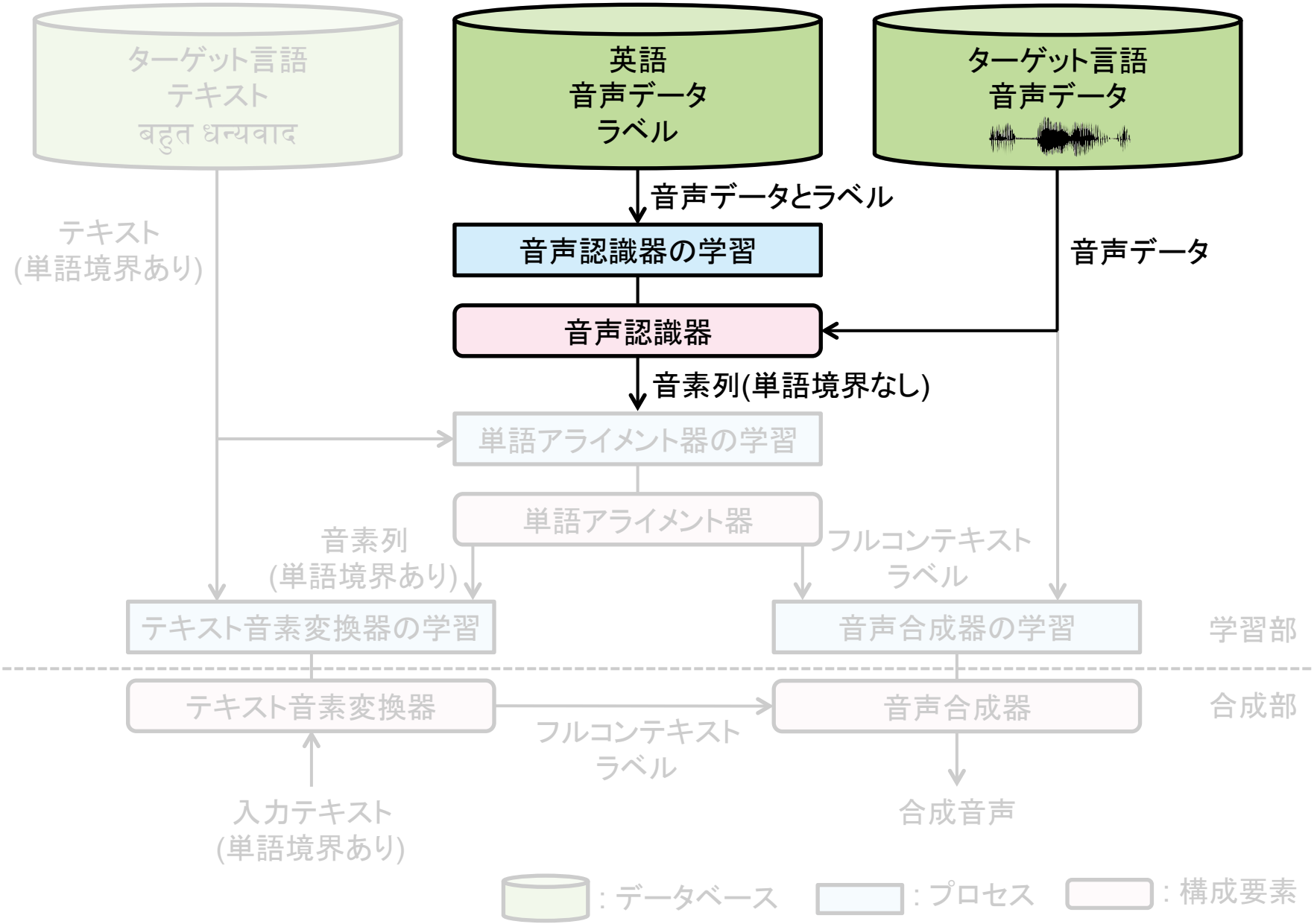
テキスト: प्रसिद्ध कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …

音声データとテキストのみからTTSシステム構築を目指す

システムの概要


ターゲット言語
テキスト
बहुत धन्यवाद
英語
音声データ
ラベル
ターゲット言語
音声データ
音声データとラベル
テキスト
(単語境界あり)
音声認識器の学習
音声データ
音声認識器
音素列(単語境界なし)
単語アライメント器の学習
単語アライメント器
音素列
(単語境界あり)
フルコンテキスト
ラベル
テキスト音素変換器の学習
音声合成器の学習
学習部
テキスト音素変換器
音声合成器
合成部
フルコンテキスト
ラベル
入力テキスト
(単語境界あり)
合成音声
: データベース
: プロセス
: 構成要素

システムの概要


ターゲット言語
テキスト
बहुत धन्यवाद
英語
音声データ
ラベル
ターゲット言語
音声データ
音声データとラベル
音声認識器の学習
音声データ
音声認識器
音素列(単語境界なし)
テキスト
(単語境界あり)
単語アライメント器の学習
単語アライメント器
音素列
(単語境界あり)
フルコンテキスト
ラベル
テキスト音素変換器の学習
音声合成器の学習
学習部
テキスト音素変換器
音声合成器
合成部
フルコンテキスト
ラベル
入力テキスト
(単語境界あり)
合成音声
: データベース
: プロセス
: 構成要素

音声認識器(SR)

- 音声データに対応する音素列
  - 別言語(英語)の不特定話者音声認識器(SISR)を利用
  - SISRの音素列より特定話者音声認識器(SDSR)を学習
  - 音素列の推定とSDSRの学習を繰り返す

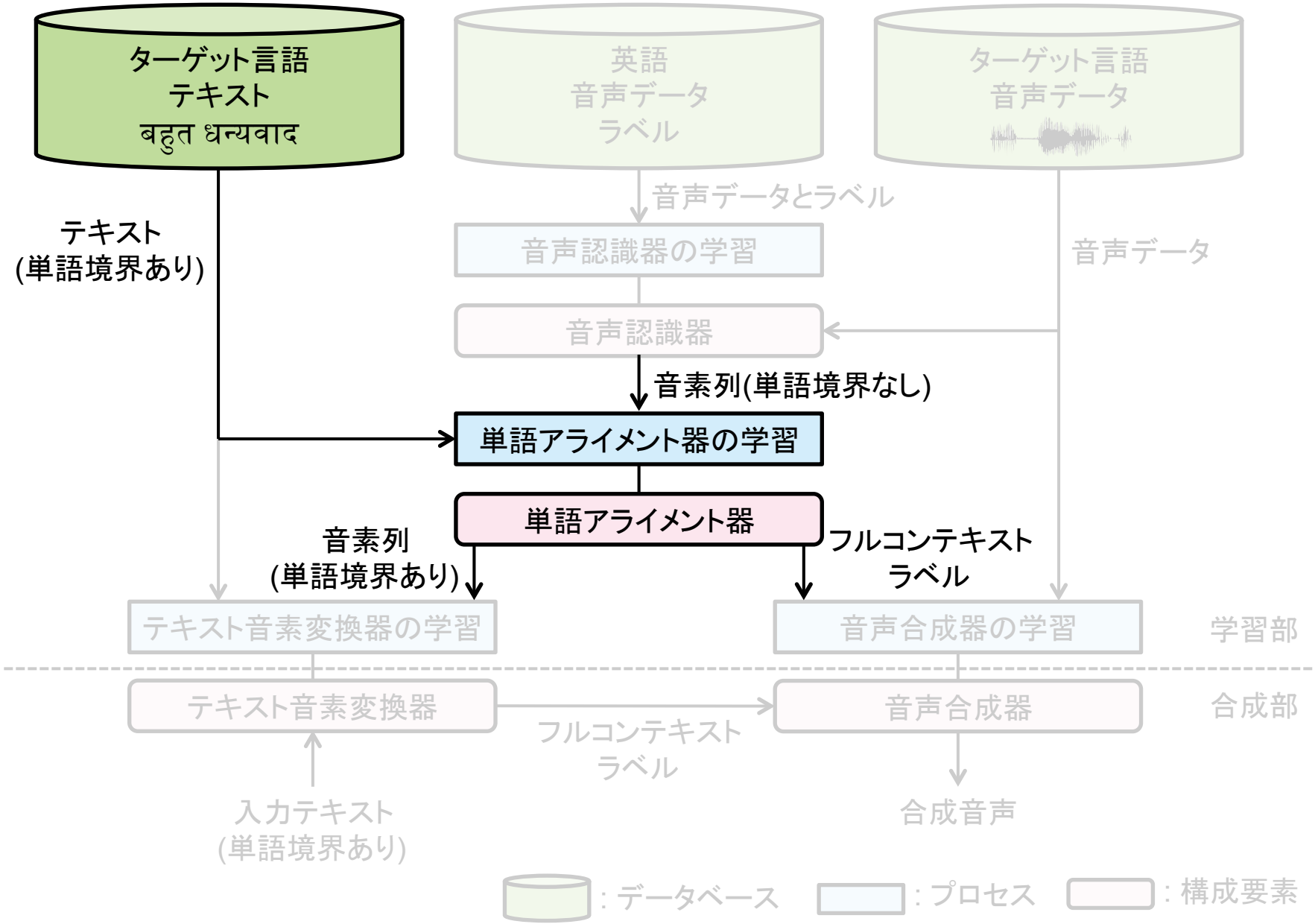
- 音素継続長を考慮した音素列
  - 音声合成器では音素継続長をモデル化
  - 通常の音声認識器では音素継続長を考慮した推定は困難
  - 隠れセミマルコフモデル(HSMM)による音素アライメント尤度を用いた音素列の選択

発音情報未知言語におけるTTSシステム構築

テキスト: प्रसिद्ध कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …

音声データに対応した音素列を取得

システムの概要


ターゲット言語
テキスト
बहुत धन्यवाद
英語
音声データ
ラベル
ターゲット言語
音声データ
音声データとラベル
テキスト
(単語境界あり)
音声認識器の学習
音声データ
音声認識器
音素列(単語境界なし)
単語アライメント器の学習
単語アライメント器
音素列
(単語境界あり)
フルコンテキスト
ラベル
テキスト音素変換器の学習
音声合成器の学習
学習部
テキスト音素変換器
音声合成器
合成部
フルコンテキスト
ラベル
入力テキスト
(単語境界あり)
合成音声
: データベース
: プロセス
: 構成要素

単語アライメント器(WA)

- 単語境界
  - 音声認識結果の音素列には単語境界は存在しない
  - 単語単位のG2P構築に単語境界は必要
  - 単語境界の情報は音声合成器のコンテキストとして有用

- Joint multigramモデルによる単語アライメント器
  - 単語境界を含むテキストと音素列の対からモデルを学習
  - テキストと音素列の対応関係を推定
  - 音声認識結果のポーズは単語境界であるという制約
  - Viterbiデコーディングにより単語境界を推定

発音情報未知言語におけるTTSシステム構築

音声データ:

SISR出力: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1出力: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2出力: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …

N-best: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

⋮

HSMM選択: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

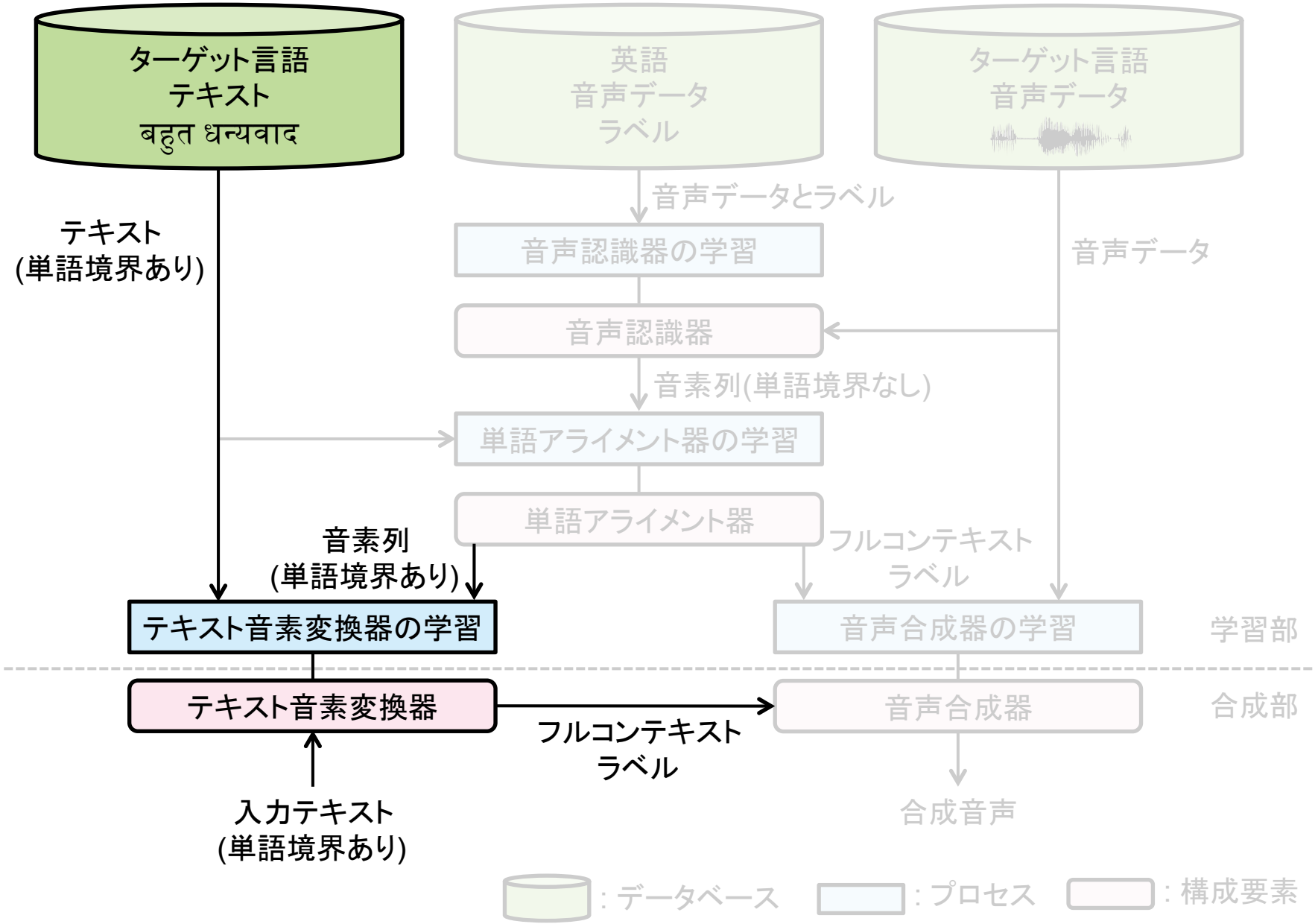
WA出力:	sil r ah s ih d r uh	g ah b iy d	ah d hh ih t ae	sil …
テキスト:	प्रसिद्ध	कबीर	अध्येता,	…

WA学習

WA

音素列の単語境界を取得

システムの概要


ターゲット言語
テキスト
बहुत धन्यवाद
英語
音声データ
ラベル
ターゲット言語
音声データ
音声データとラベル
テキスト
(単語境界あり)
音声認識器の学習
音声データ
音声認識器
音素列(単語境界なし)
単語アライメント器の学習
単語アライメント器
音素列
(単語境界あり)
フルコンテキスト
ラベル
テキスト音素変換器の学習
音声合成器の学習
学習部
テキスト音素変換器
音声合成器
合成部
フルコンテキスト
ラベル
入力テキスト
(単語境界あり)
合成音声
: データベース
: プロセス
: 構成要素

テキスト音素変換器(G2P)

- 任意のテキストを音素列に変換する必要
  - 発音情報が未知の言語では辞書構築は困難
  - 辞書の代わりにテキスト音素変換器を構築

- Joint multigramモデルによるテキスト音素変換器 [Bisani, et at.; ’08]
  - 単語単位のテキストと音素列の対からモデルを学習
  - Viterbiデコーディングにより音素列を推定

発音情報未知言語におけるTTSシステム構築

音声データ:

SISR出力: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1出力: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

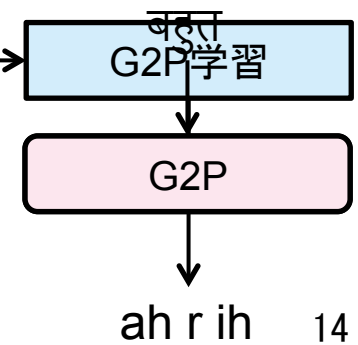
SDSR2出力: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …

N-best: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

⋮

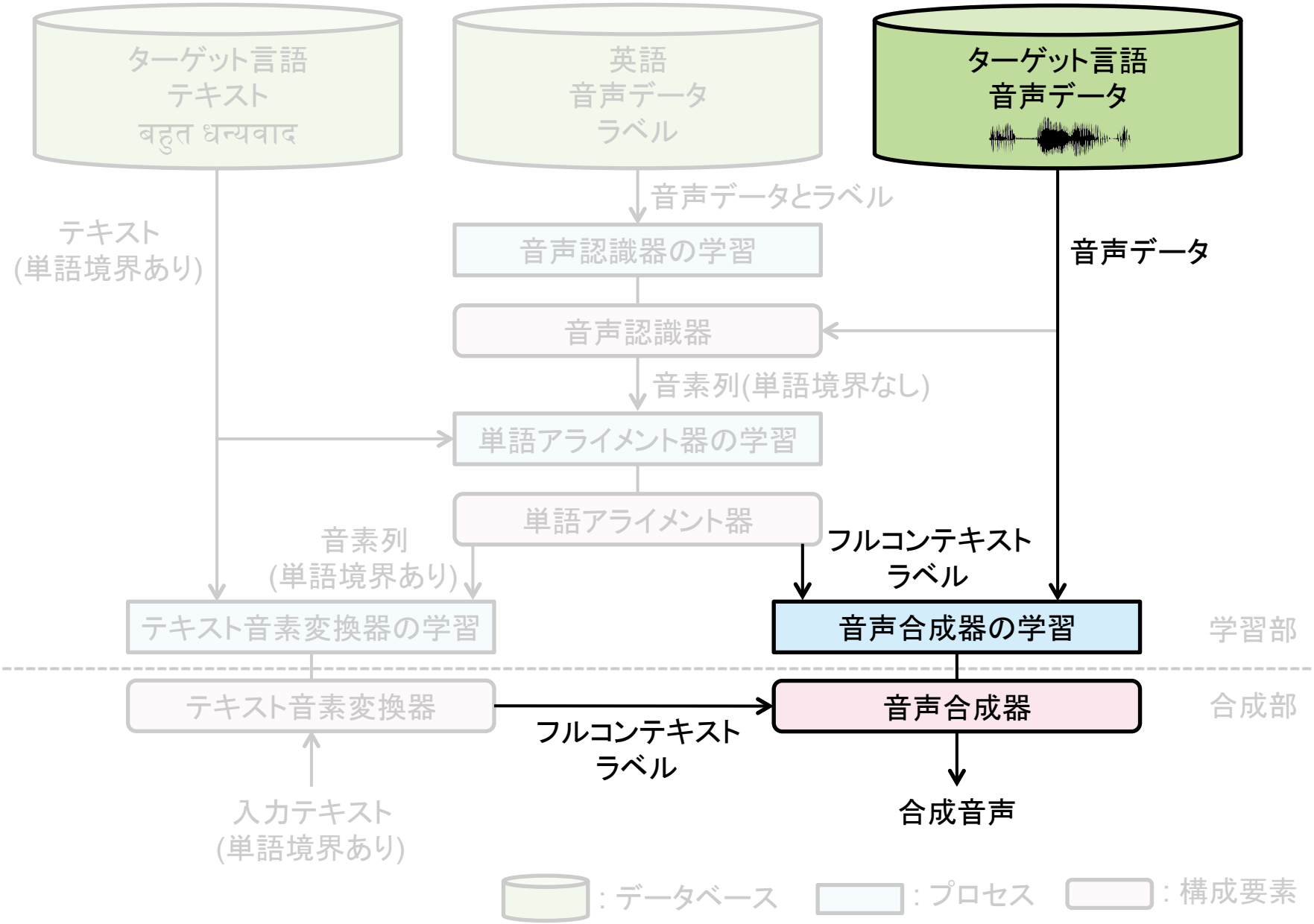
HSMM選択: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

WA出力:	sil r ah s ih d r uh	g ah b iy d	ah d hh ih t ae	sil …
テキスト:	प्रसिद्ध	कबीर	अध्येता,	…



任意のテキストを音素列に変換可能

システムの概要

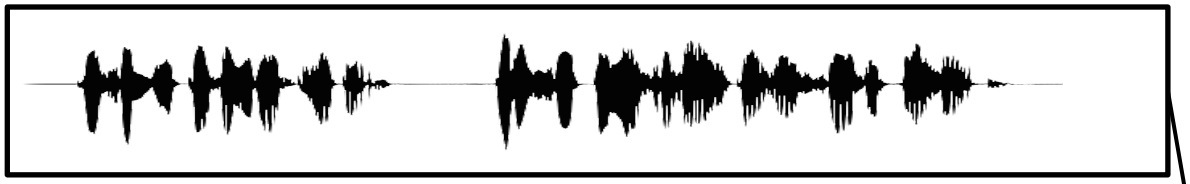

ターゲット言語
テキスト
बहुत धन्यवाद
英語
音声データ
ラベル
ターゲット言語
音声データ
音声データとラベル
テキスト
(単語境界あり)
音声認識器の学習
音声データ
音声認識器
音素列(単語境界なし)
単語アライメント器の学習
単語アライメント器
音素列
(単語境界あり)
フルコンテキスト
ラベル
テキスト音素変換器の学習
音声合成器の学習
学習部
テキスト音素変換器
音声合成器
合成部
フルコンテキスト
ラベル
入力テキスト
(単語境界あり)
合成音声
: データベース
: プロセス
: 構成要素

音声合成器(SS)

- 音声合成器にはコンテキストの設計が重要
- コンテキストの設計
  - クインフォン
    - 音声認識器の結果を利用
  - 音節
    - *C***V*を仮定 (*C*: 子音, *V*: 母音, *C**: 0回以上の子音の繰返し)
  - 単語
    - 単語アライメント器の結果を利用
  - 句
    - 音声認識結果のポーズを利用
  - 文章

発音情報未知言語におけるTTSシステム構築

音声データ:



SISR出力: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1出力: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2出力: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …

N-best: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

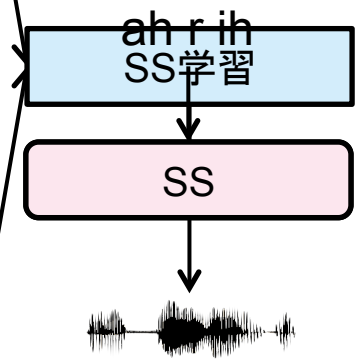
⋮

HSMM選択: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

WA出力:

sil r ah s ih d r uh	g ah b iy d	ah d hh ih t ae	sil …

テキスト: प्रसिद्ध कबीर अध्येता, …



フルコンテキストラベルから音声を合成可能

BLIZZARD CHALLENGE

- インド語合成音声

2014			
Assamese			
Gujarati			
Hindi			
Rajasthani			
Tamil			
Telugu			

2015			
Bengali			
Hindi			
Malayalam			
Marathi			
Tamil			
Telugu			

日本人による評価実験が困難

実験条件 (1/2)

- ターゲット言語
  - 日本語
    - ひらがなのみを仮定
    - 文節を分かち書き境界と仮定
    - 例: てれびげえむや ぱそこんで げえむお して あそぶ
- 音声認識器

英語データベース	TIMIT, WSJ0, WSJ1
日本語データベース	ATR音声データベースセットB MHT話者
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
モデル構造	3状態スキップなしleft-to-right HMM 32(ポーズ)と16(ポーズ以外)混合のGMM
音素列推定と適応の繰り返し回数	3

実験条件 (2/2)

音声合成器

日本語データベース	ATR音声データベースセットB MHT話者
サンプリング周波数	16 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

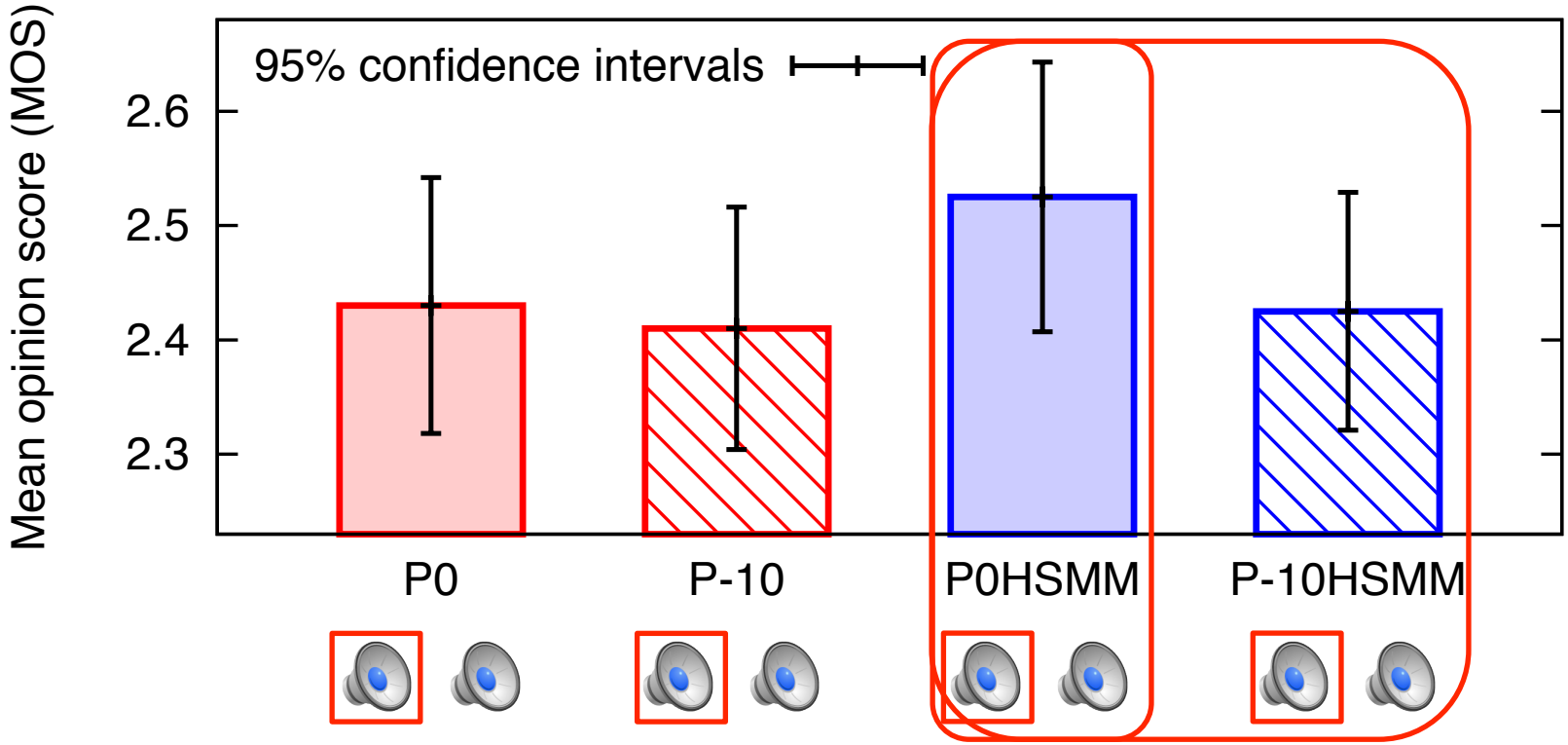
実験内容

評価手法	自然性に関する5段階MOS試験
被験者数	10人
評価文章	学習データには用いていない53文からランダムに20文

実験結果 (1/2)

- 挿入ペナルティとHSMMによる音素列選択の影響

挿入ペナルティ \ HSMM	なし(最上位仮説)	上位50仮説から選択
0	P0	P0HSMM
−10	P−10	P−10HSMM

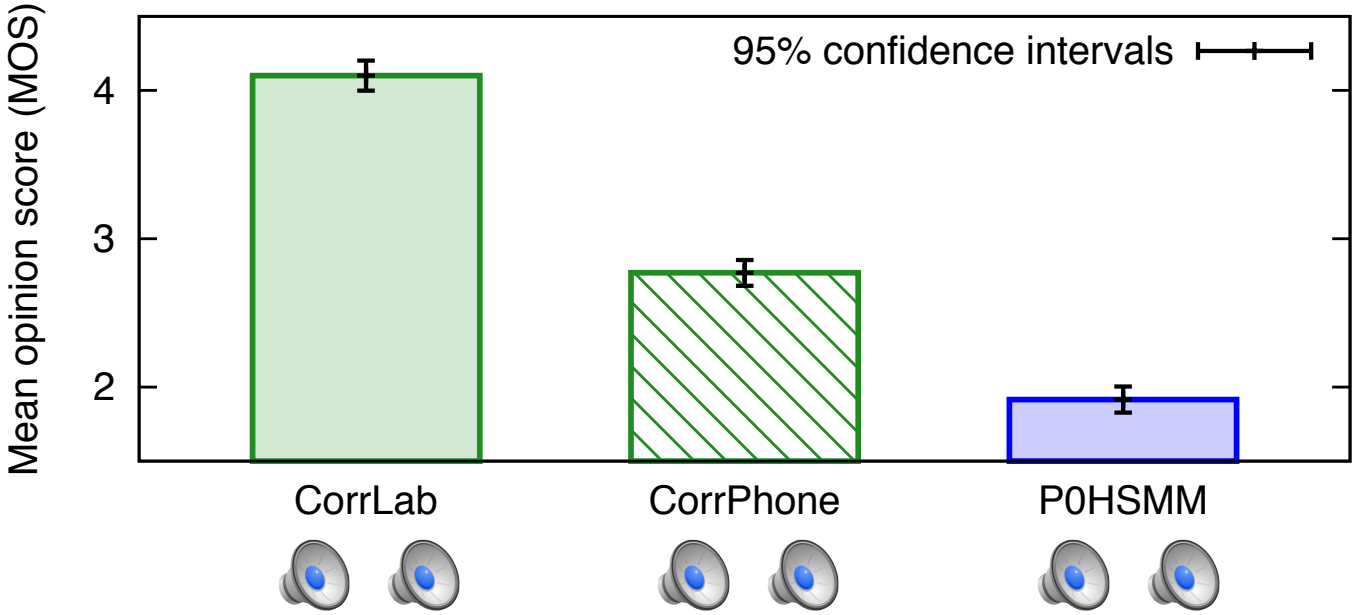


適切な挿入ペナルティの設定が必要

実験結果 (2/2)

- 正解情報を用いた手法との比較

	音素セット	学習時		合成時
		音素列	単語境界やアクセント等	入力
CorrLab	日本語	正解	あり	正解フルコンテキスト
CorrPhone	日本語	正解	なし	テキスト
P0HSMM	英語	推定	なし	テキスト



むすび

- 発音情報が未知の言語のTTSシステム構築
  - 専門的な知識を利用せずにTTSシステムを自動構築
  - 主観評価実験
    - HSMMの音素アライメント尤度による音素列の選択は有効
    - 適切な挿入ペナルティの設定が必要
    - 自然性を高めるためには各構成要素の精度を高める必要

- 今後の課題
  - 挿入ペナルティなどの設定基準の検討
  - 国際音声記号による多言語不特定話者音声認識器の構築
  - 音声データに基づいた教師なし音素セット決定法の検討