The NITECH HMM-based text-to-speech system for the Blizzard Challenge 2015

Kei Sawada, Kei Hashimoto,
Keiichiro Oura, and Keiichi Tokuda

Nagoya Institute of Technology (NITECH)

Blizzard Challenge 2015 workshop on Sep 11, 2015

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

Background

- Text-to-speech (TTS) systems
  - System to synthesize speech for arbitrary text
  - TTS have been used widely in various applications
    - Car navigation, smart phone, spoken dialogue system, etc.
  - Demand for TTS systems has increased
    - High speech quality, multilingual language, speaking styles, etc.

- Multilingual language TTS systems
  - Thousands of languages exist in the world
  - To establish a framework that can be applied to build TTS system of any target languages
    ⇒ One goal of speech synthesis research
  - Require a special knowledge of a target language

Focus on automatically constructing TTS of any languages

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

Blizzard Challenge 2015 tasks

- Blizzard Challenge [Black, et al.; '05]
  - Blizzard Challenge was started in order to better understand and compare research techniques
- Speech synthesis for six Indian languages
  - Bengali, Hindi, Malayalam, Marathi, Tamil, Telugu
- Hub task
  - Build TTS systems in each Indian language
  - Provide speech data and corresponding to text
- Spoke task
  - Build a multilingual (polyglot) TTS system (Indian and English)
  - Training data for the Spoke task is same as for the Hub task
  - Sample input text (Hindi and English):

    Stanford वैज्ञानिकों द्वारा विकसित नयी aluminium battery,
    केवल एक minute में cellphone को charge कर सकती है.

Blizzard Challenge 2015 tasks

Waveform: 

Need to construct TTS system from only speech data and corresponding to text

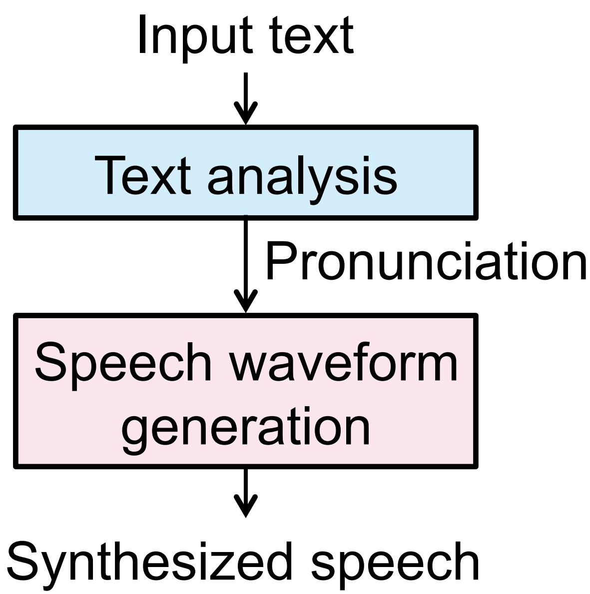
Text: प्रसद्द कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

TTS system components

- Text analysis part
  - Estimate pronunciation of an input text
  - Using a lexicon containing pronunciation information
  - High language-dependency



- Speech waveform generation part
  - Speech waveforms are generated from pronunciation info.
  - Unit-selection [Hunt, et al. '96]
  - Statistical parametric speech synthesis (SPSS)
    - Hidden Markov model (HMM) [Tokuda, et al.; '00]
    - Deep neural network (DNN) [Zen, et al.; '13]
    - Low language-dependency

TTS system construction

- Steps of TTS system construction
  - Acoustic features are modeled by using phoneme information

Target language DB

Speech data

Text

बहुत धन्यवाद

Step 1: Definition of a phoneset
Step 2: Construction of a lexicon or G2P
Step 3: Design of contextual factors
Step 4: Preparation of label sequences

Require a special knowledge of the target language
⇒ Huge cost for someone not familiar with the target language

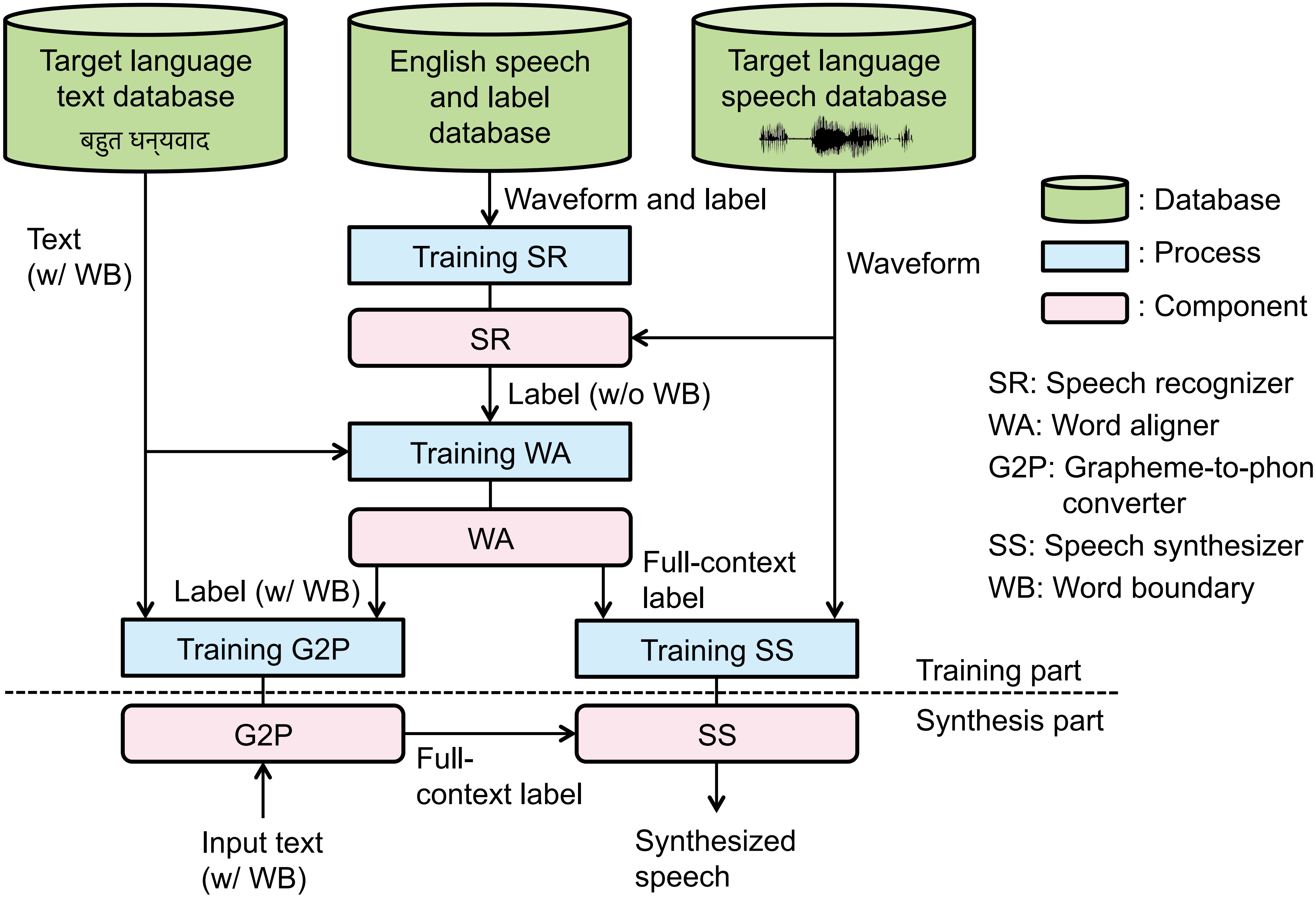
TTS system

Investigate a framework to automatic TTS construction in any unknown-pronunciation languages

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

System overview



SR: Speech recognizer

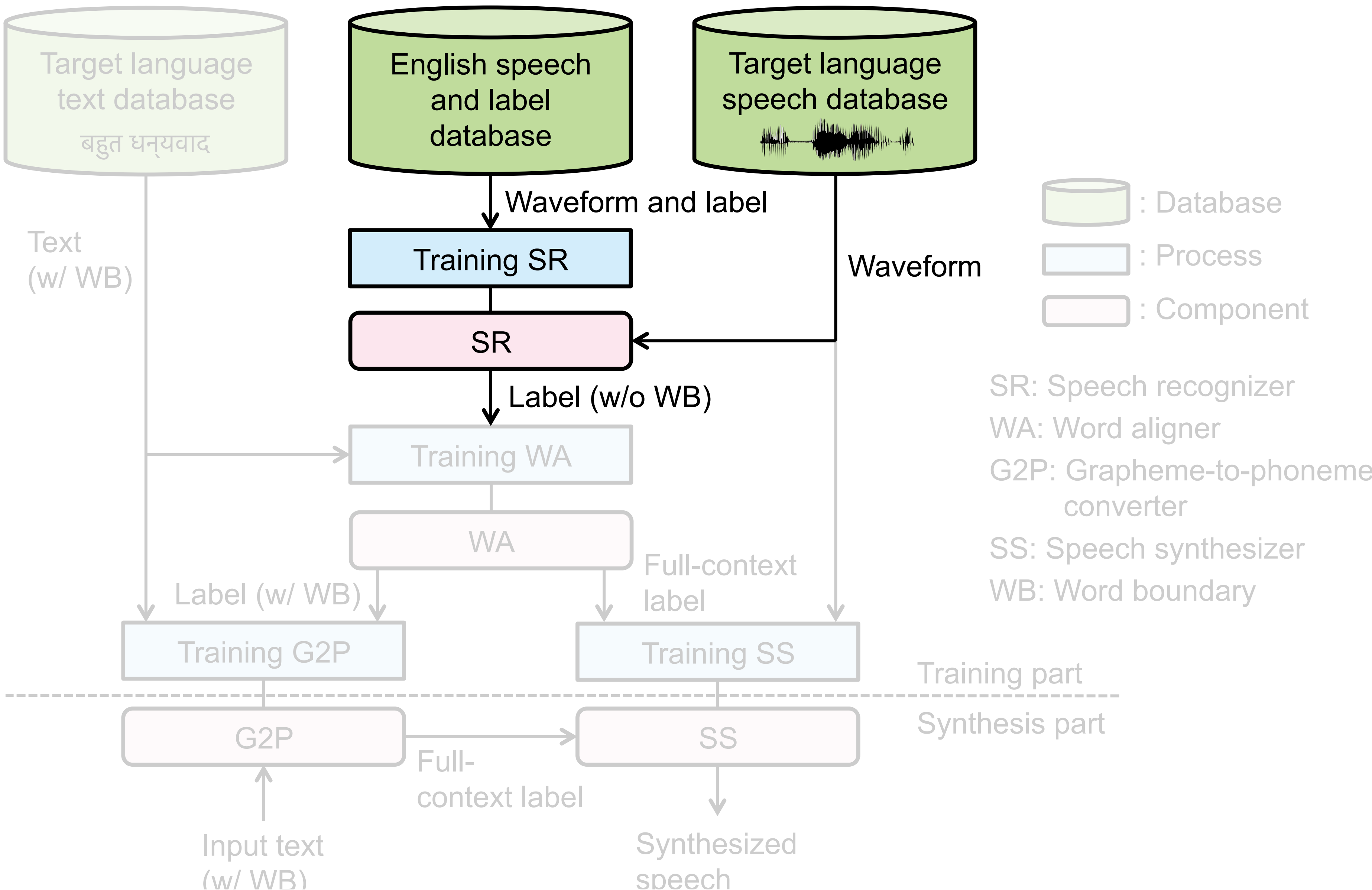
WA: Word aligner

G2P: Grapheme-to-phoneme converter

SS: Speech synthesizer

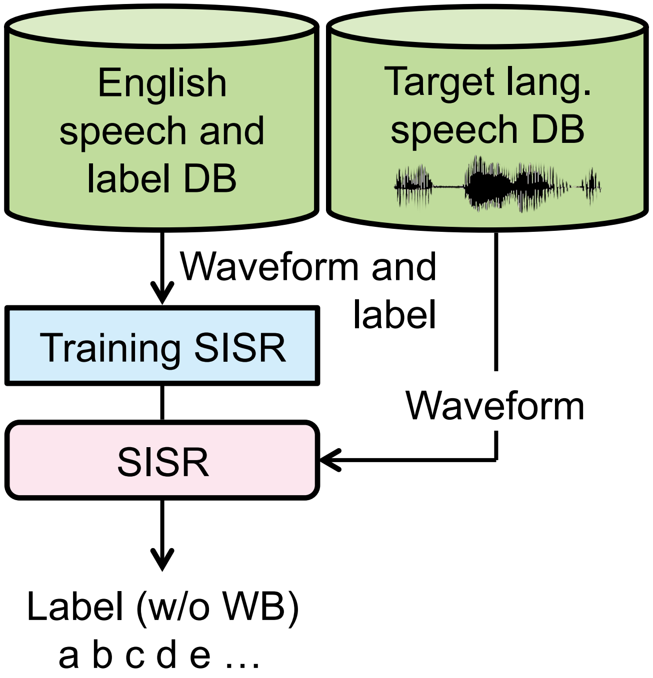
WB: Word boundary

System overview



Speech recognizer (SR) (1/3)

- Phoneme sequence corresponding to speech data
  - Speech recognition is carried out by using speaker-independent SR (SISR) of the other language, e.g., English
  - Triphone recognizer
  - Phoneset of SISR is used as phonset of target language



Construction flow

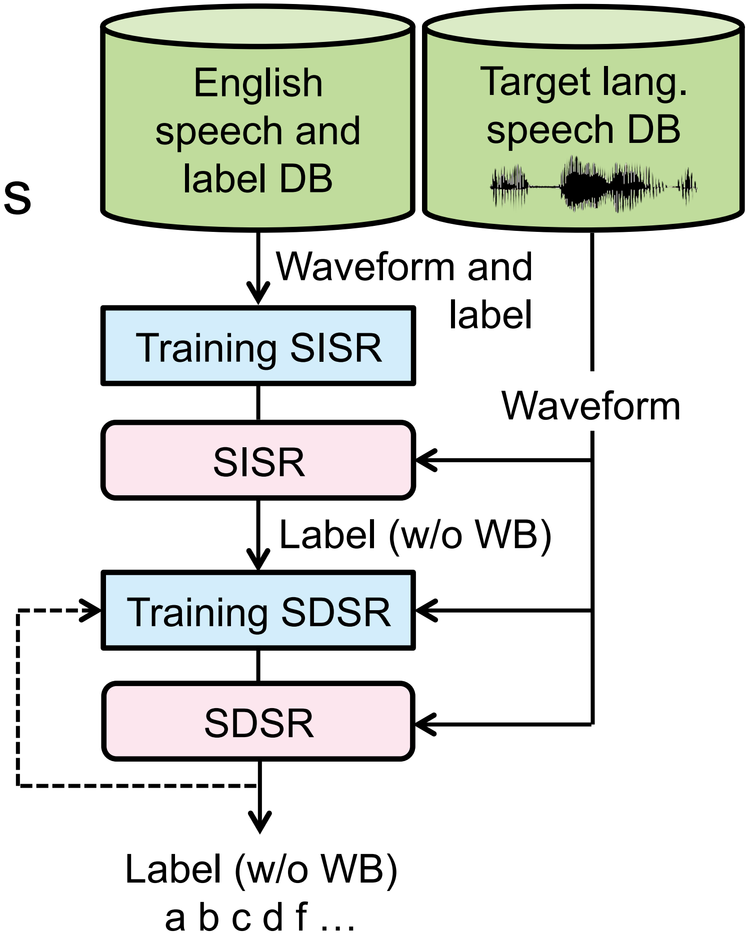
Waveform:

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

Text: प्रसिद्द कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …

Speech recognizer (SR) (2/3)

- More accurate phoneme sequence
  - Speaker-dependent SR (SDSR) is constructed from phoneme sequences obtained by the SISR.
  - Estimation of phoneme sequences and training of SDSR are iterated
- Unsupervised training
  - Phoneme sequences obtained by the SISR is initial value
  - Data-driven training by repetition



Construction flow

Waveform:

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

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

Training SDSR2

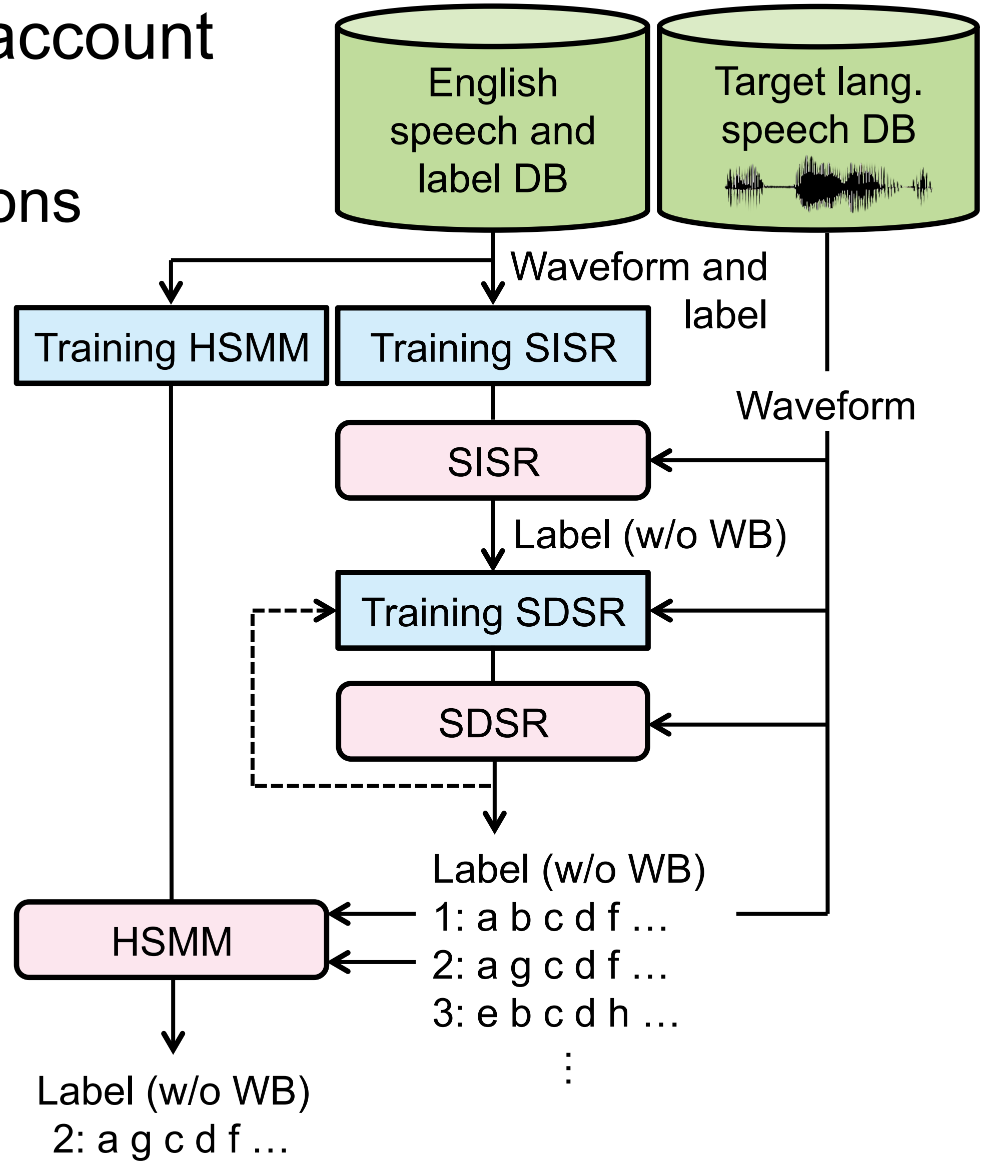
SDSR2

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

Text: प्रसदि्द कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …

Speech recognizer (SR) (3/3)

- Phoneme sequence taking account of phoneme duration
  - Modeling of phoneme durations is important for an SS
  - SR is difficult to consider phoneme duration
  - Phoneme sequence is selected using alignment likelihood of HSMM
  - Phoneme sequence with the highest alignment likelihood is selected as phoneme sequence



Construction flow

Waveform:

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

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

SDSR2 output:

N-best

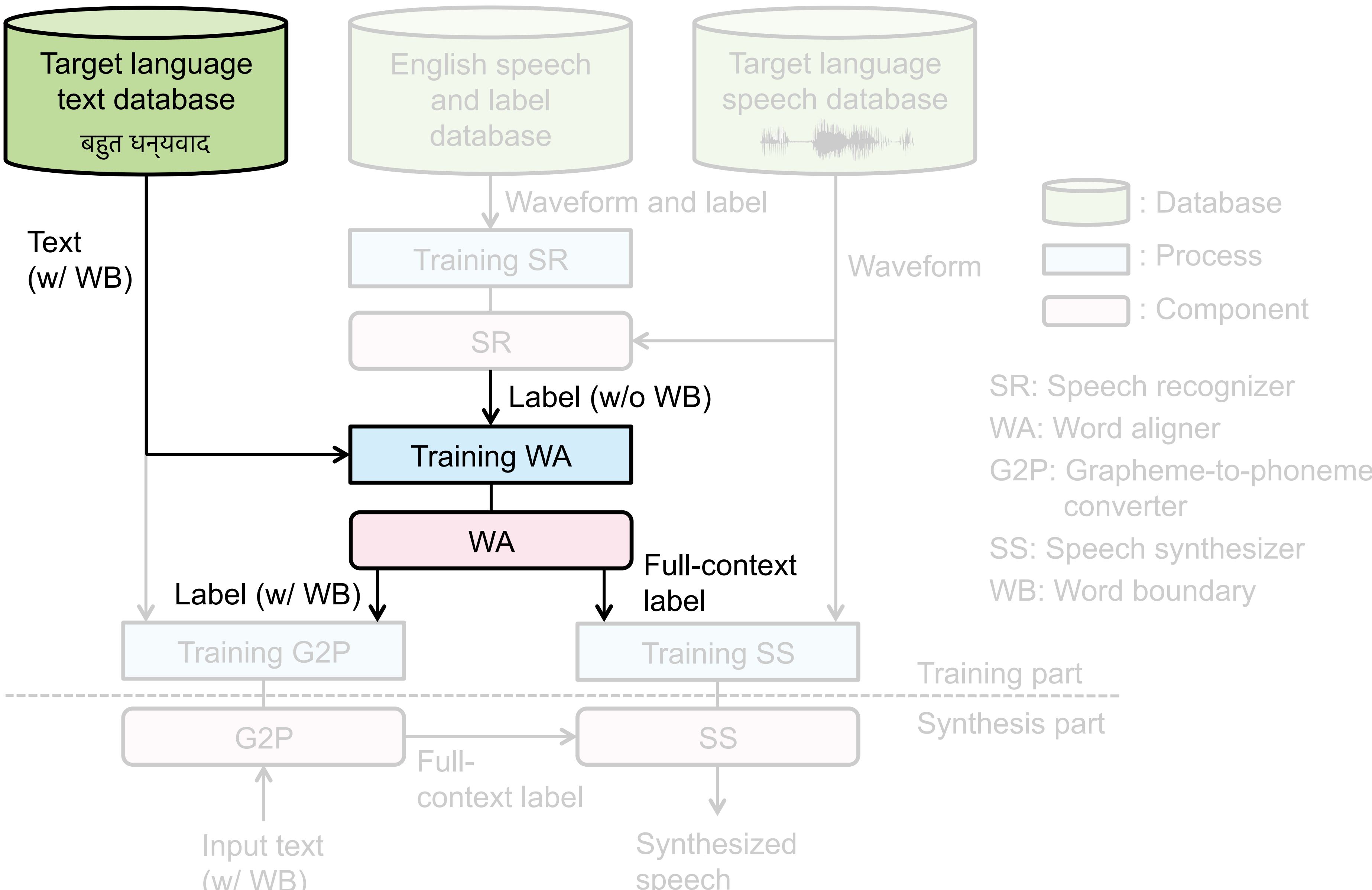
- sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …
- sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …
- ⋮

→ HSMM selection

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

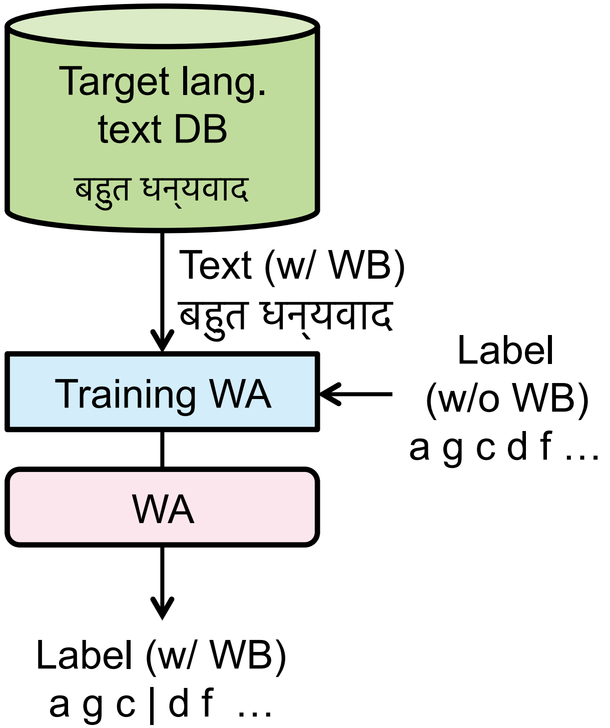
Obtain phoneme sequences of target language speech

System overview



Word aligner (WA)

- Word boundary (WB) information
  - Word-level G2P is suitable
  - WB information is useful for contextual factors of the SS
  - Phoneme sequences obtained by speech recognition does not include WB



- Joint multigram model-based WA
  - WA is constructed for estimation of WB
  - A pause of recognition results must be WB
  - Viterbi decoding

Construction flow

Waveform: 

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

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

SDSR2 output:

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

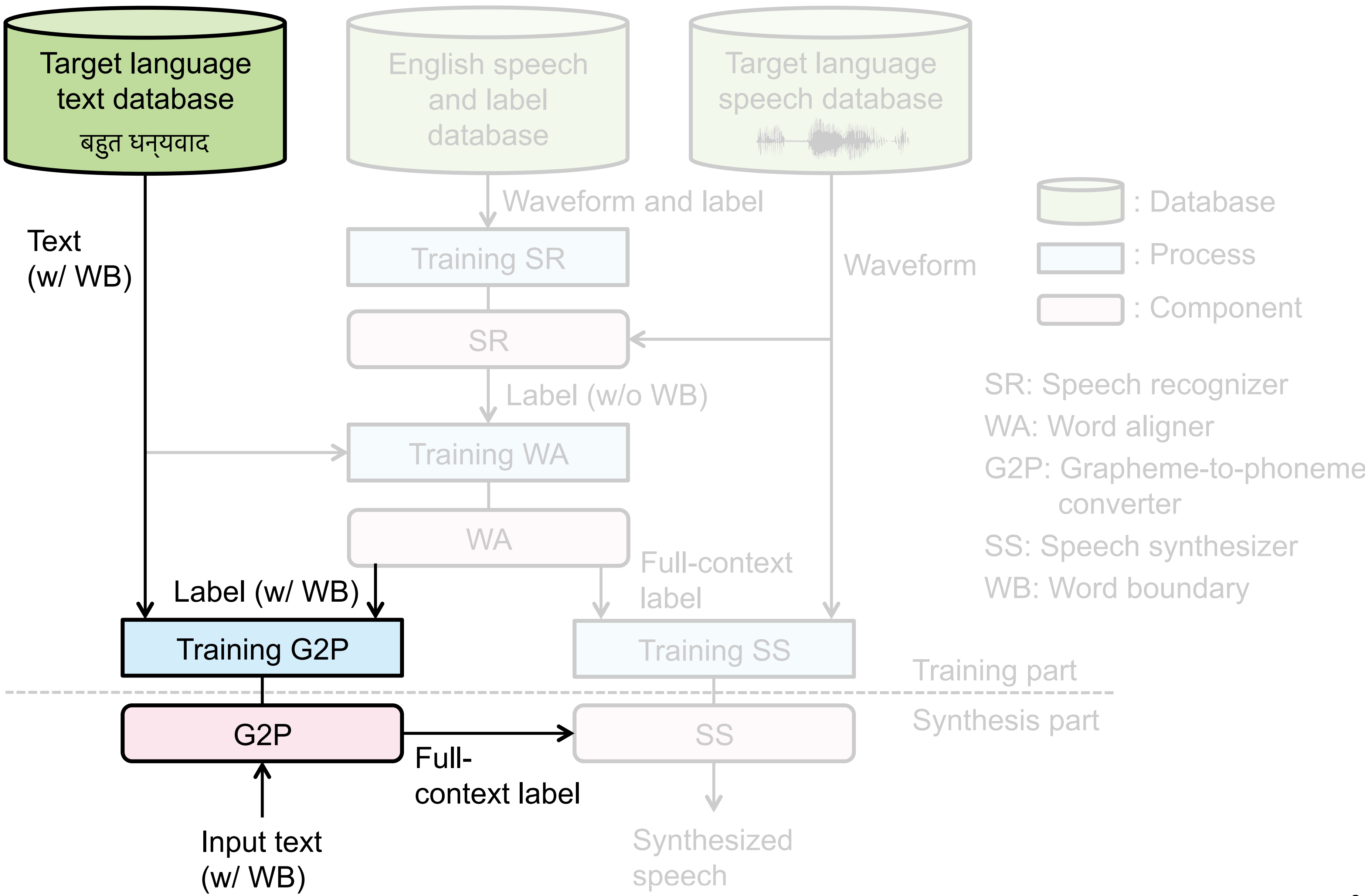
HSMM out

Obtain WB information of phoneme sequence

WA output:	sil r ah s ih d r uh	g ah b iy d	ah d hh ih t ae	sil …
Text:	प्रसिद्द	कबीर	अध्येता,	…

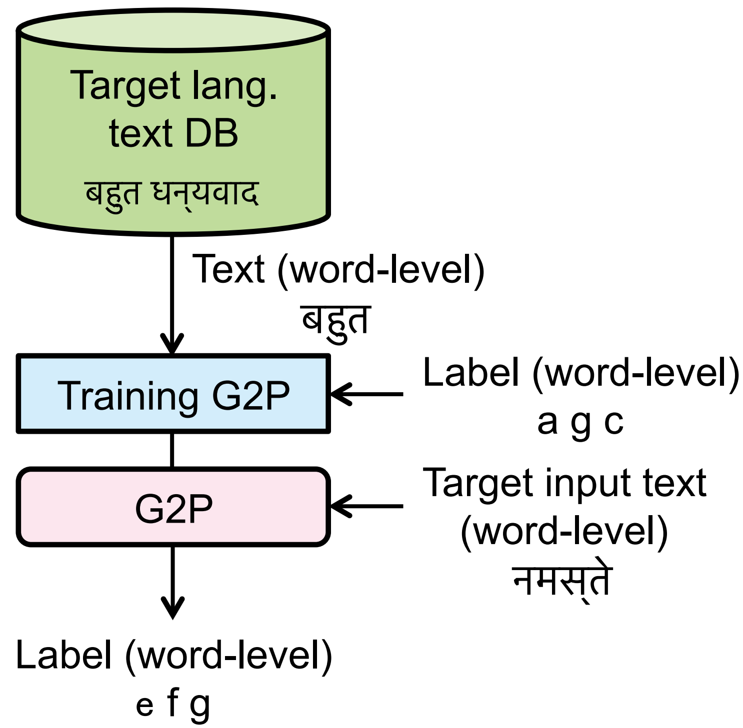
→ Training WA — WA

System overview



Grapheme-to-phoneme converter (G2P) (1/2)

- Arbitrary input text need to be converted into phoneme sequence
  - Difficult to construct a lexicon in unknown-pronunciation language
- Joint multigram model-based G2P [Bisani, et at.; ’08]
  - G2P is introduced instead of lexicon
  - Viterbi decoding



- Pause insertion
  - Comma, colon, parenthesis
  - Before or after a word that is easy to enter pause in a speech recognition result

Construction flow

Waveform:

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

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

SDSR2 output:
N-best
- sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …
- sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …
- ⋮

HSMM o

Obtain phoneme sequence of arbitrary input text

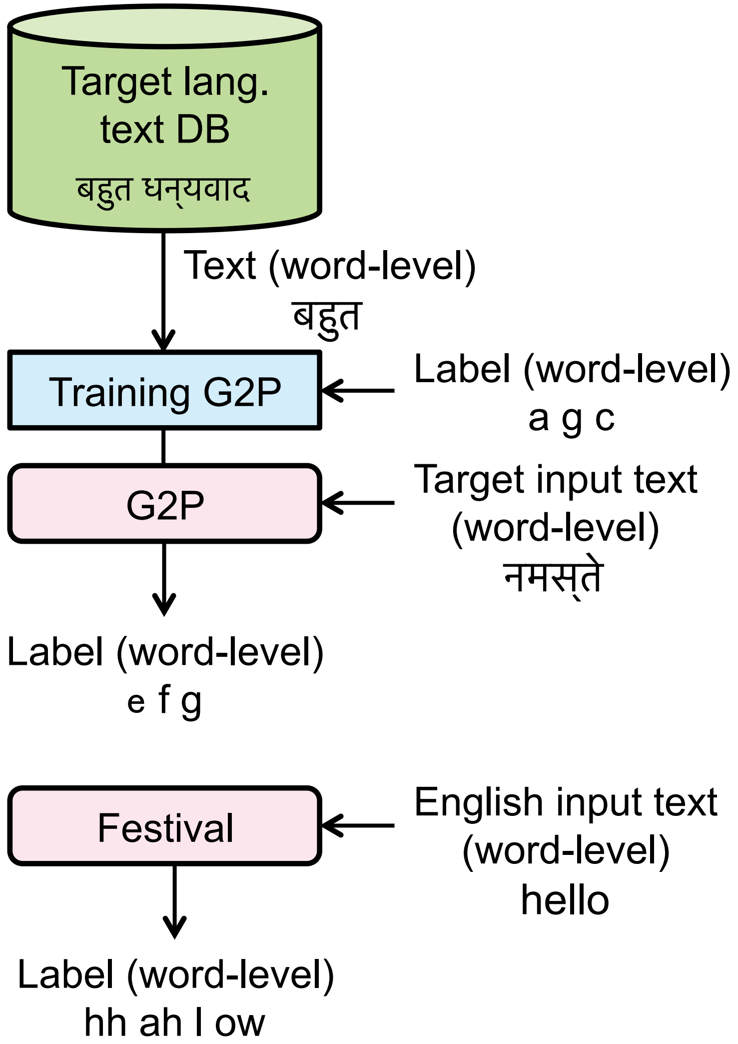
WA output:	sil r ah s ih d r uh	g ah b iy d	ah d hh ih t ae	sil …
Text:	प्रसद्दि	कबीर	अध्येता,	…

Training G2P

G2P

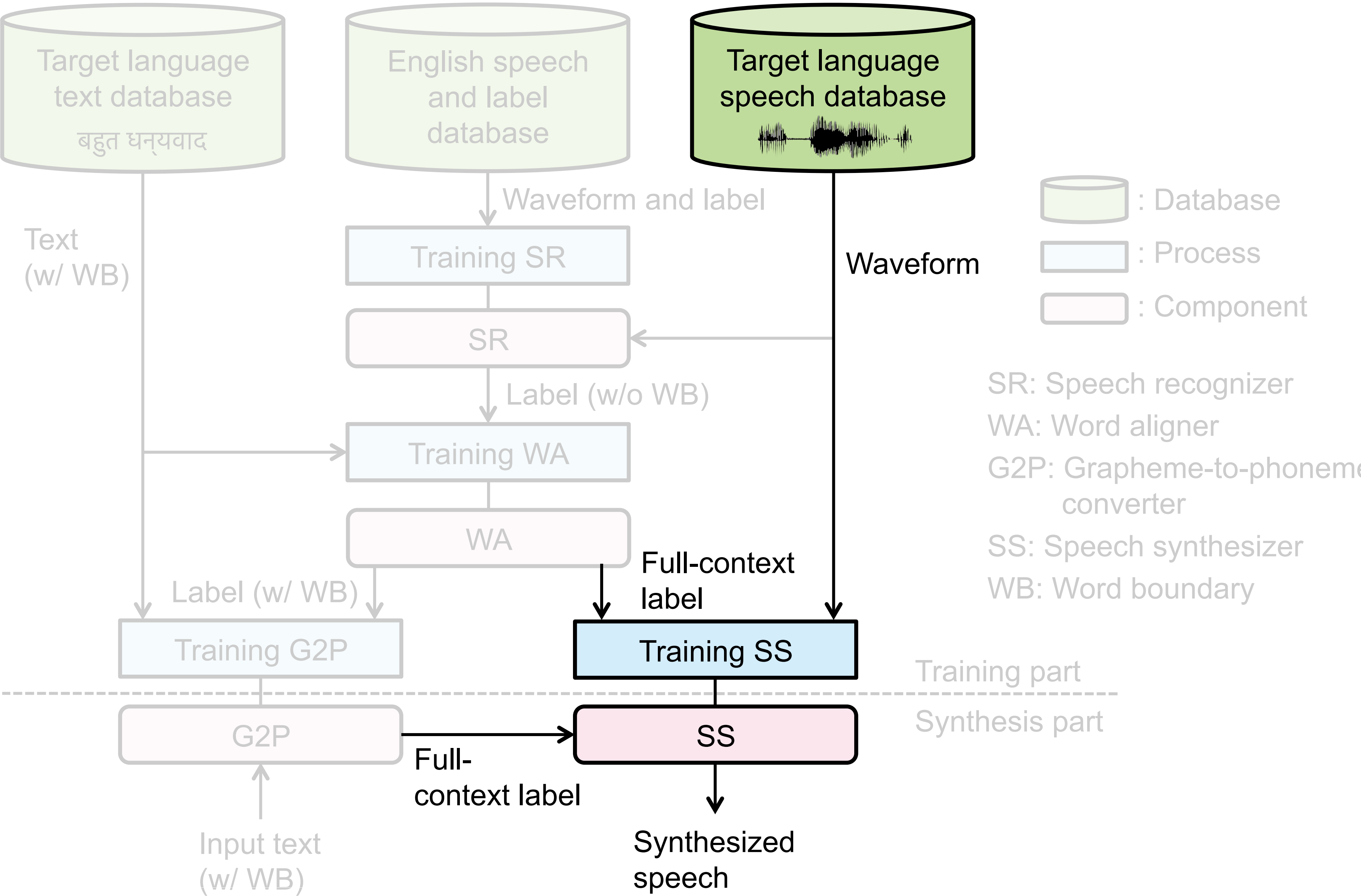
Grapheme-to-phoneme converter (G2P) (2/2)

- Multilingual speech synthesis
  - Input text includes Indian language and English
  - Phoneset of acoustic model is the same as the English SISR
  - Can synthesize Indian language and English
  - Indian language text analysis: G2P
  - English text analysis: Festival



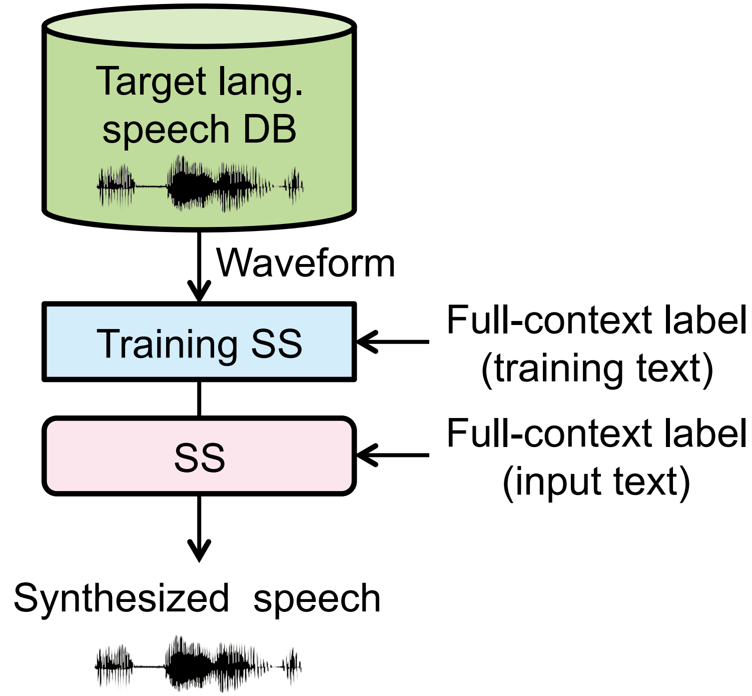
Can synthesize multilingual speech

System overview



Speech synthesizer (SS)

- Design of contextual factors
  - Quint phone
  - Syllable
    - Defined as *C*V* (*C*: consonant, *V*: vowel *C**: none or more *C*)
  - Word
  - Phrase
  - Utterance



- Base techniques
  - STRAIGHT [Kawahara, et al.; ’99], MSD [Tokuda, et al.; ’02], HSMM [Zen, et al.; ’04], GV [Toda, et al.; ’05]

Construction flow

Waveform:

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

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

SDSR2 output:
N-best
- sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …
- sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …
- ⋮

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

WA output: sil r ah s ih d r uh | g ah b iy d | ah d hh ih t ae | sil …

Training SS

SS

T

Can synthesize speech from full-context label

Strengths and weaknesses

- Strengths
  - Low language-dependency
    - Can apply to languages in which sentences written with space between words
  - Multilingual speech synthesis
    - Can synthesize target language and English
- Weaknesses
  - Pronunciation (text analysis) error
    - Pronunciation errors can occur duo to errors in SR, WA, and G2P
  - Difficult to adjust each components
    - I don’t understand what synthesized speech is saying

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

Speech recognizer conditions

English database	WSJ0, WSJ1, and TIMIT
Indian database	Six Indian language
Window	Hamming window
Frame	length: 25 ms, shift: 10 m
Feature vector	12-dimension MFCC + Δ + ΔΔ
Model structure	3-state left-to-right HMM without skip transition

	Bengali	Hindi	Malayalam	Marathi	Tamil	Telugu
Insertion penalty	−20	−40	−40	−40	−20	−10
Number of iteration	3	2	3	3	3	2

Speech synthesizer conditions

Indian database	Six Indian language
Sampling rate	16.0 kHz
Window	F0-adaptive Gaussian window
Frame shift	5 ms
Feature vector	39-dim. STRAIGHT mel-cepstrum, log F0, 19-dmi. aperiodicity measure $+ \Delta + \Delta\Delta$
Model structure	5-state left-to-right MSD-HSMM without skip transition

	Bengali	Hindi	Malayalam	Marathi	Tamil	Telugu
Number of sentences	1284	1690	1269	1178	1440	2461
Time	1h58m27s	3h57m59s	1h58m9s	2h6m7s	4h9m28s	4h11m34s

Evaluation conditions

Evaluation criteria	Intelligibility (word error rate), MOS of speaker similarity, MOS of naturalness
System A	Natural speech
System G	NITECH system

RD	Read text
SUS	Semantically unpredictable sentences
ML	Multilingual sentences (Indian and English)

	Bengali	Hindi	Malayalam	Marathi	Tamil	Telugu
Number of listeners	48	69	72	69	70	70

Word error rate (SUS)

	Bengali	Hindi	Malayalam	Marathi	Tamil	Telugu
	WER	WER	WER	WER	WER	WER
A	43	42	59	35	29	82
B	52	49	64	35	57	70
C	100	34	50	26	49	46
D	57	25	78	30	49	60
E	69	40	47	52	69	54
F	66	23	74	41	67	59
G	76	30	98	69	47	75
H	55	24	46	35	57	46
I	100	31	73	-	50	62
J	61	40	52	21	60	57

Important to properly adjusts SR, WA, G2P for each language

MOS of speaker similarity

	Bengali			Hindi			Malayalam			Marathi			Tamil			Telugu		
	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML
A	4.5	4.7	4.6	4.5	4.4	4.5	4.6	4.2	4.2	4.4	4.3	4.3	4.6	4.6	4.2	4.5	3.3	3.8
B	2.5	3.7	1.8	2.6	3.5	1.8	1.8	2.1	2.6	2.3	2.7	2.2	1.8	1.9	2.2	2.1	2.1	2.1
C	2.2	2.2	2.1	2.4	2.0	2.2	2.3	2.1	2.2	2.3	1.9	1.6	2.2	2.8	2.8	1.3	1.4	1.5
D	2.2	3.0	2.8	2.7	2.2	2.0	2.1	2.2	2.1	3.0	3.4	3.1	1.9	2.0	1.7	2.0	2.6	2.3
E	3.7	3.3	4.1	2.9	2.8	3.5	2.9	2.6	3.2	2.9	3.4	2.7	2.7	2.6	3.3	2.9	2.5	2.5
F	1.7	2.6	2.1	4.3	3.9	3.1	2.3	2.3	2.6	3.0	2.8	2.7	2.7	2.5	2.6	2.5	2.0	2.2
G	3.1	3.1	2.2	2.8	2.9	2.2	2.3	2.1	2.0	2.5	2.2	2.1	2.3	3.1	2.4	3.1	2.2	1.4
H	2.4	2.5	-	2.7	2.6	-	2.0	2.2	-	2.1	2.4	-	2.6	2.7	-	2.4	3.4	-
I	2.9	2.3	-	3.5	3.3	-	3.0	3.2	-	-	-	-	3.6	3.4	-	4.2	1.9	-
J	2.7	2.8	-	3.3	2.2	-	2.0	2.9	-	2.8	2.4	-	2.6	2.3	-	2.7	2.0	-

The high MOS even though WER is the high rate (76%)
⇒ High WER or pronunciation errors scarcely affect MOS

MOS of naturalness

	Bengali			Hindi			Malayalam			Marathi			Tamil			Telugu		
	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML	RD	SUS	ML
A	4.7	4.6	4.7	4.7	4.4	4.7	4.3	4.3	4.4	4.6	4.5	4.8	4.7	4.6	4.7	4.8	4.5	4.8
B	2.2	2.7	1.8	3.2	2.6	1.8	1.6	1.9	1.9	2.7	2.5	2.2	2.2	2.2	2.2	1.9	1.8	2.0
C	2.9	3.1	2.6	3.5	3.3	3.2	2.6	2.8	2.4	2.5	2.7	2.6	2.8	3.3	2.9	2.6	2.5	2.5
D	3.0	3.0	2.6	2.8	2.7	2.3	2.3	2.4	2.2	3.0	2.9	2.6	2.6	2.6	2.3	2.1	2.5	2.6
E	3.4	2.7	3.8	2.6	3.0	3.2	2.3	2.7	3.6	3.0	3.3	3.4	2.5	3.0	4.0	2.8	2.7	2.9
F	2.0	1.8	1.6	3.9	3.9	2.9	2.9	2.5	2.7	3.2	3.2	2.9	3.6	3.2	3.3	3.5	2.5	2.6
G	2.5	2.8	2.2	2.3	2.4	2.0	1.7	2.0	1.9	2.2	2.2	2.1	2.4	2.3	2.6	2.1	2.1	2.4
H	2.6	2.6	-	2.8	3.0	-	2.1	2.1	-	2.9	2.7	-	3.0	3.7	-	3.0	3.0	-
I	2.7	2.1	-	2.8	3.2	-	2.7	2.9	-	-	-	-	3.2	3.0	-	2.9	2.1	-
J	2.6	2.5	-	3.3	3.1	-	2.9	2.3	-	3.3	2.9	-	2.7	3.0	-	3.5	2.7	-

The low MOS even though WER is the lowest rate (47%)
⇒ Even a little word pronunciation error often affects MOS

Speech samples

	Bengali	Hindi	Malayalam	Marathi	Tamil	Telugu
RD						
SUS						
ML						

Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions

Conclusions

- TTS system developed for the Blizzard Challenge 2015
  - Investigated automatic TTS system construction in an unknown-pronunciation language
  - Enabled a target language and multilingual TTS construction
  - Achieved a high score if the SR, WA, and G2P component can properly construct

- Future work
  - Investigation of a construction criteria
  - Construction of the multilingual SISR using the IPA
  - Investigation of phoneset determination approaches based on speech data