Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014

Kei Sawada, Shinji Takaki, Kei Hashimoto,
Keiichiro Oura, and Keiichi Tokuda
Nagoya Institute of Technology (NITECH)

Blizzard Challenge 2014 Workshop on Sep 19, 2014

Outline

Background

Blizzard Challenge 2014 rules

System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

Experiments

Conclusions

Background

Text-to-speech (TTS) system

- TTS have been used widely in various applications
  - Car navigation, mobile phone, spoken dialogue, etc.
- Main components of TTS system
  - Text analysis: lexicon
  - Speech waveform generation: unit-selection [Hunt, et al.], hidden Markov model (HMM) [Tokuda[1], et al.], deep neural network [Zen[1], et al.]

Blizzard Challenge [Black, et al.]

- Blizzard Challenge was started in order to better understand and compare research techniques

NITECH has participated using HMM-base TTS

Outline

Background

Blizzard Challenge 2014 rules

System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

Experiments

Conclusions

Blizzard Challenge 2014 rules

TTS systems of six Indian languages

- Assamese, Gujarati, Hindi, Rajasthani, Tamil, Telugu

Hub task (IH1)

- Build one voice TTS system in each Indian language
- Provided speech data and corresponding text

Spoke task (IH2)

- Build a multilingual TTS system (Indian and English)
- Training data for this task was same as for Hub task
- Sample input text (Hindi and English):

उन्हें 10 दन तक rehab करना होगा और उसके बाद उनका fitness test लया जाएगा

Difficulty in TTS system building

Phoneset of target Indian language doesn’t exist

- Use a speech recognizer of English
  - Obtain label sequences of target Indian language
  - Also useful for multilingual speech synthesis

Label sequence doesn’t include word breaking info.

- Use multigram word aligner
  - Obtain word breaking information of label sequence

Lexicon of target Indian language doesn’t exist

- Use joint multigram grapheme-to-phoneme converter
  - Obtain label sequences of given input text

Outline

Background

Blizzard Challenge 2014 rules

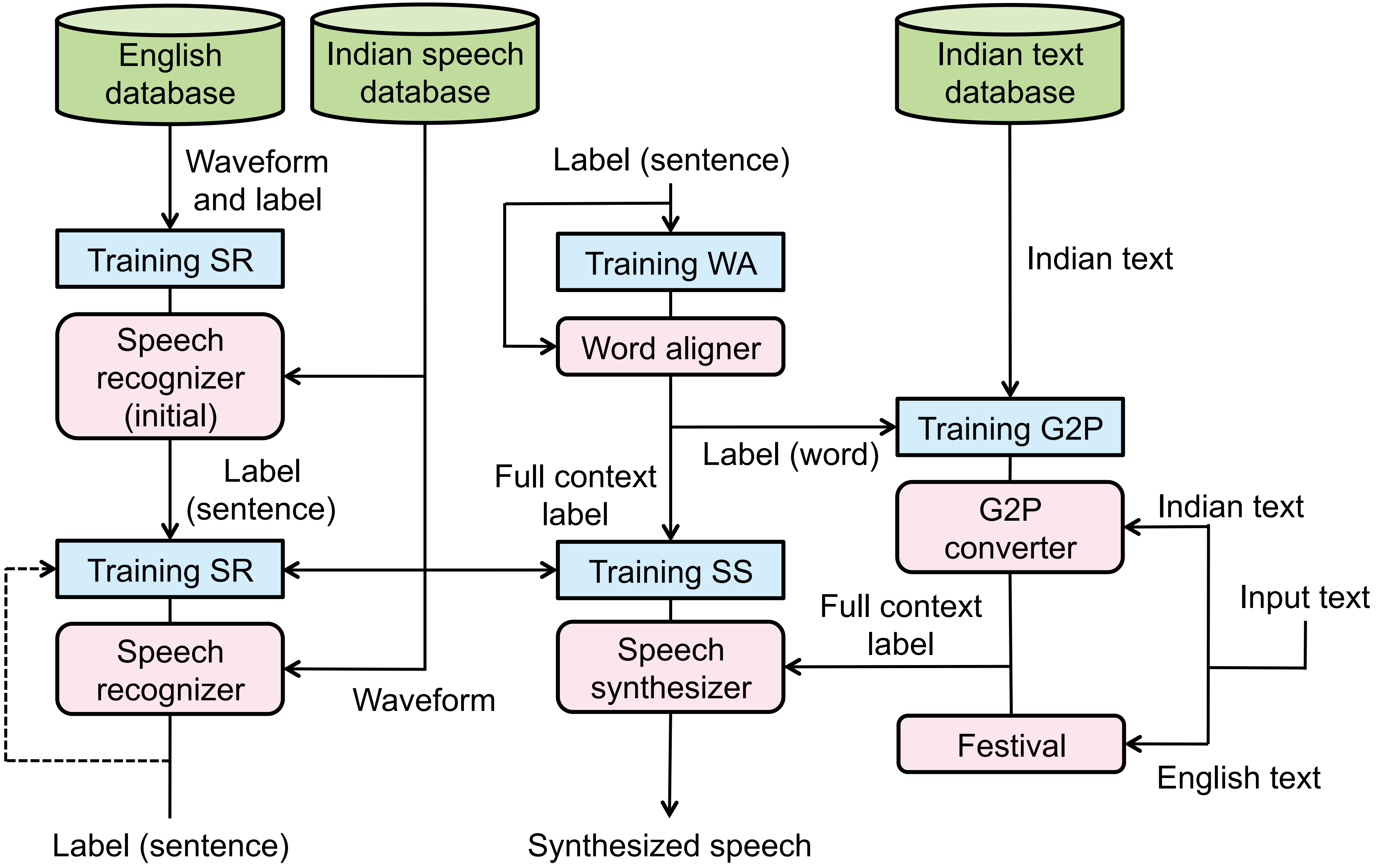
System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

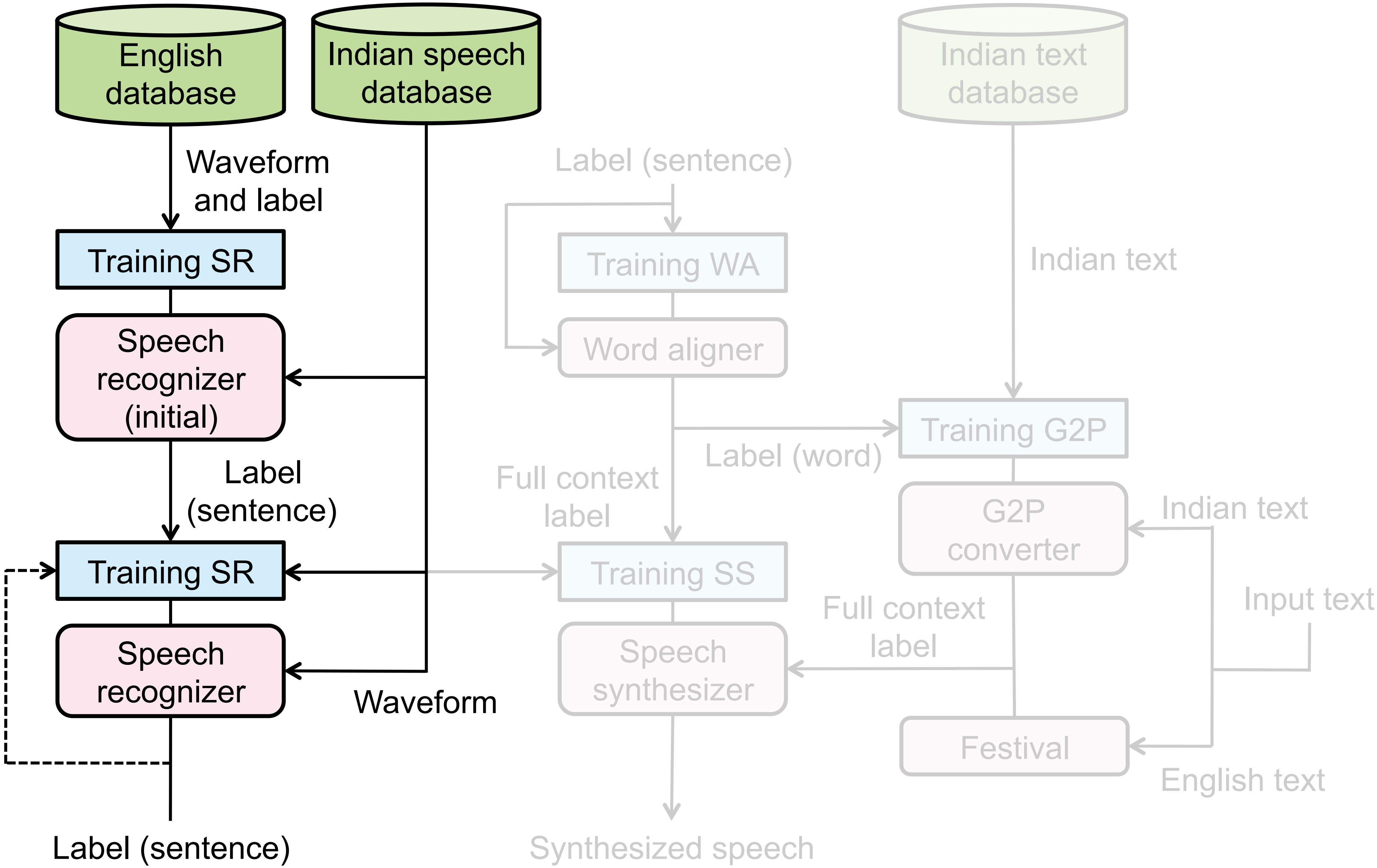
Experiments

Conclusions

System overview



System overview

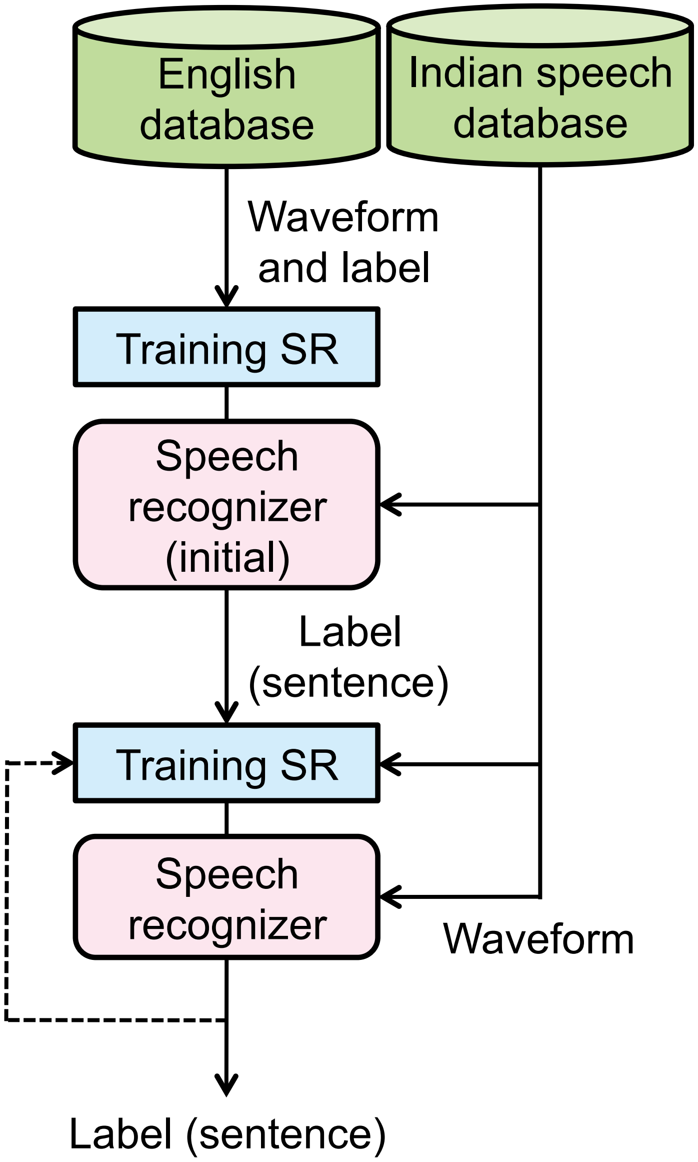


Speech recognizer (SR)

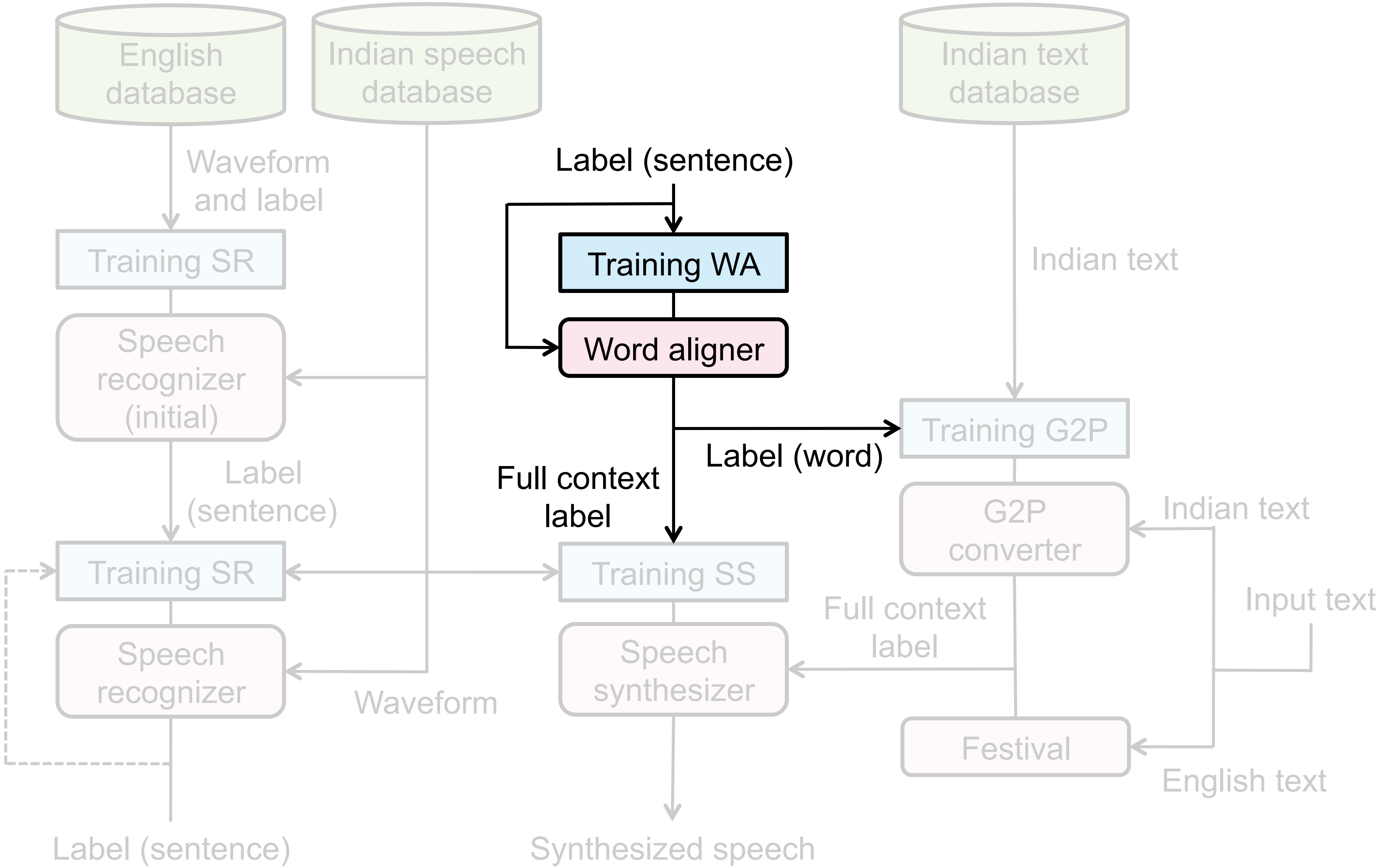
Speech recognizer

- Initial SR is built by using English
  - WSJ0, WSJ1, and TIMIT databases are used
- SR is built by using recognized label sequences
- To obtain high accuracy SR, SR is re-trained

⇒ Obtain label sequences of target Indian language speech



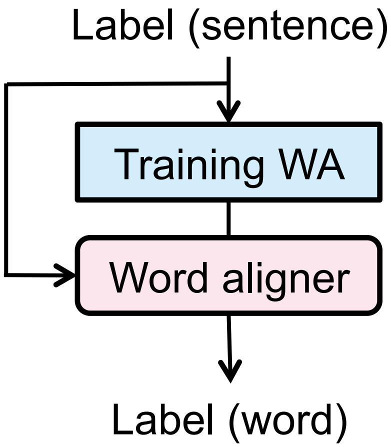
System overview



Word aligner (WA)

Word breaking information

- Word breaking information is required for full context labels of speech synthesis
- Word-level G2P converter is required

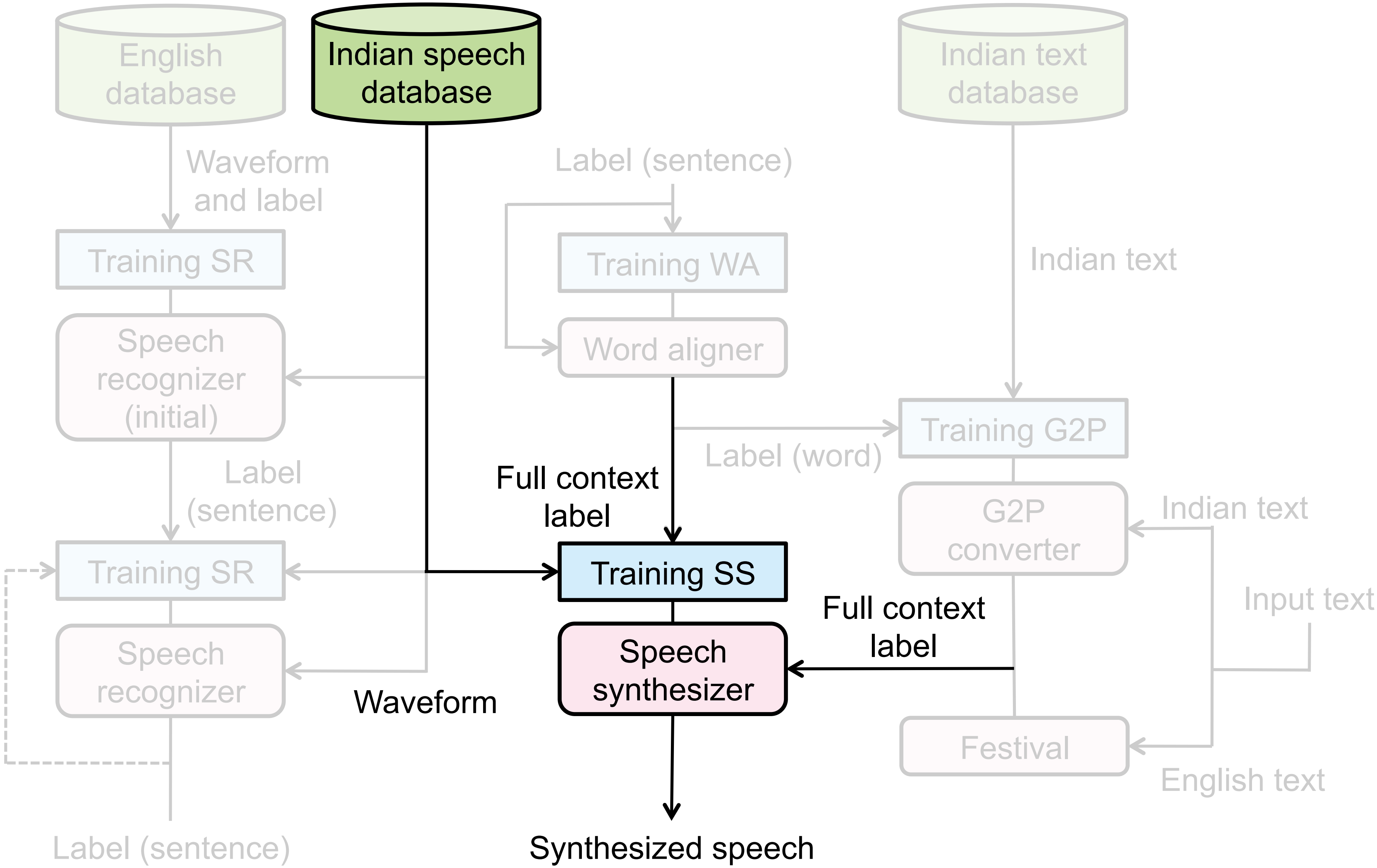


Multigram word aligner [Deligne, et al.]

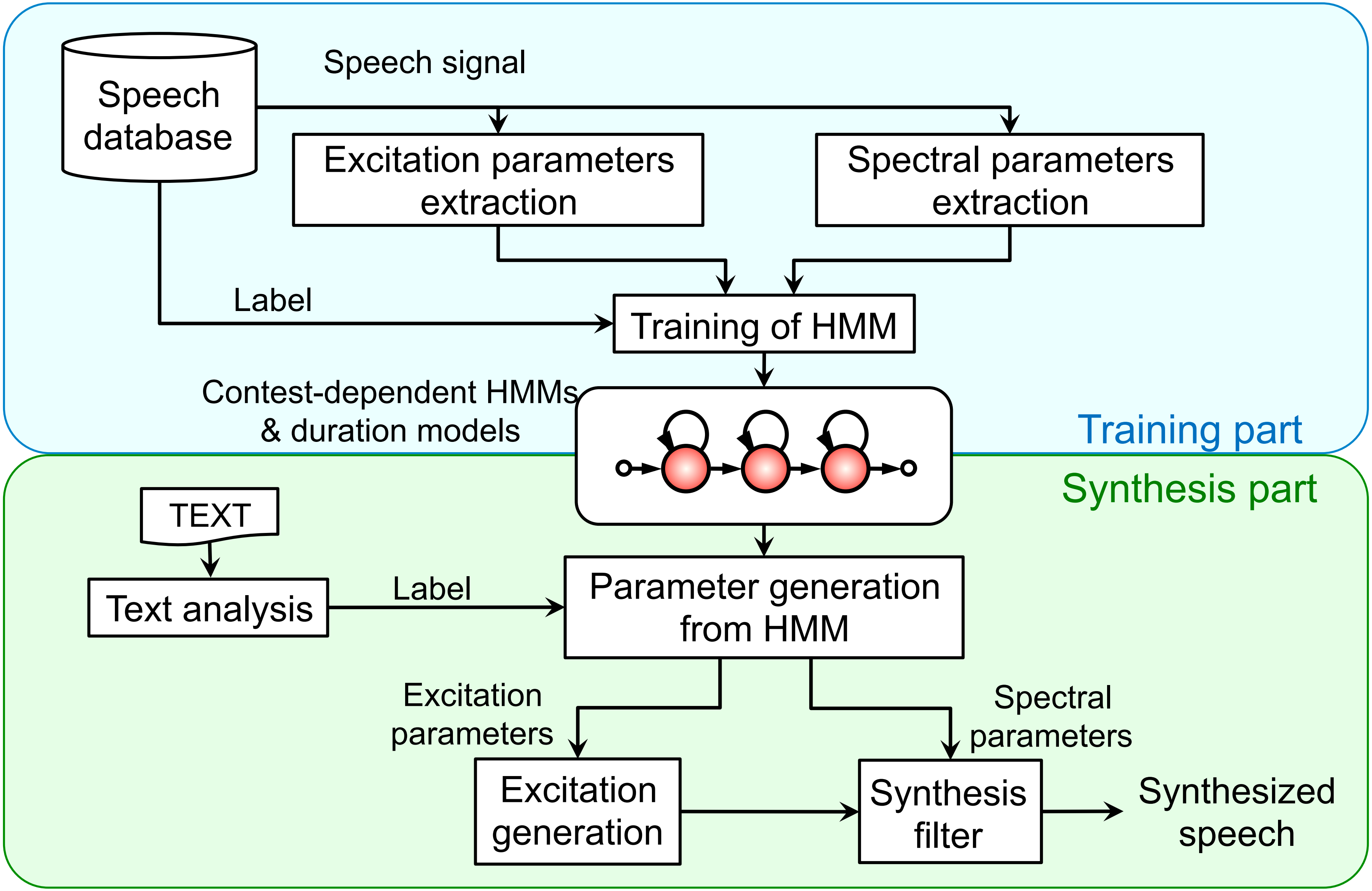
- Multigram models are estimated by using EM algorithm
- Word alignment is obtained by applying Viterbi algorithm

⇒ Obtain word breaking information of label sequences

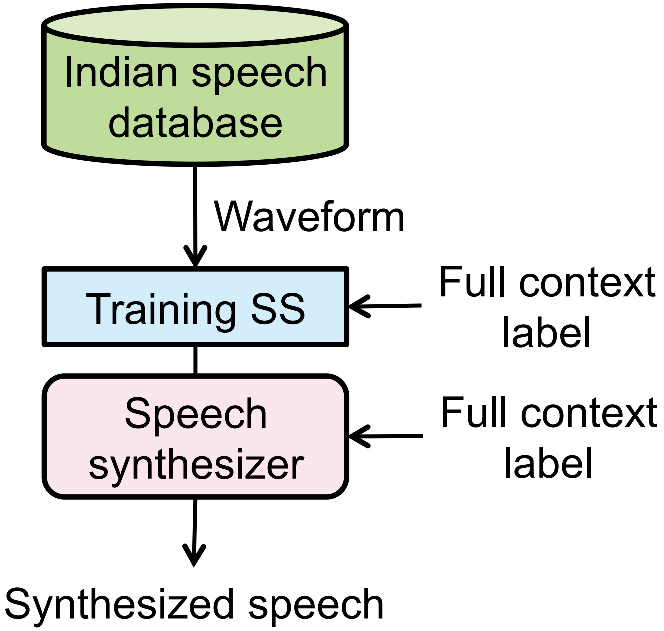
System overview



Speech synthesizer (SS)



Base techniques of SS



HSMM [Zen^2, et al.]

- HMM with explicit state duration probability distribution

MSD [Tokuda^2, et al.]

- Output distributions consist of continuous dist. and discrete dist.

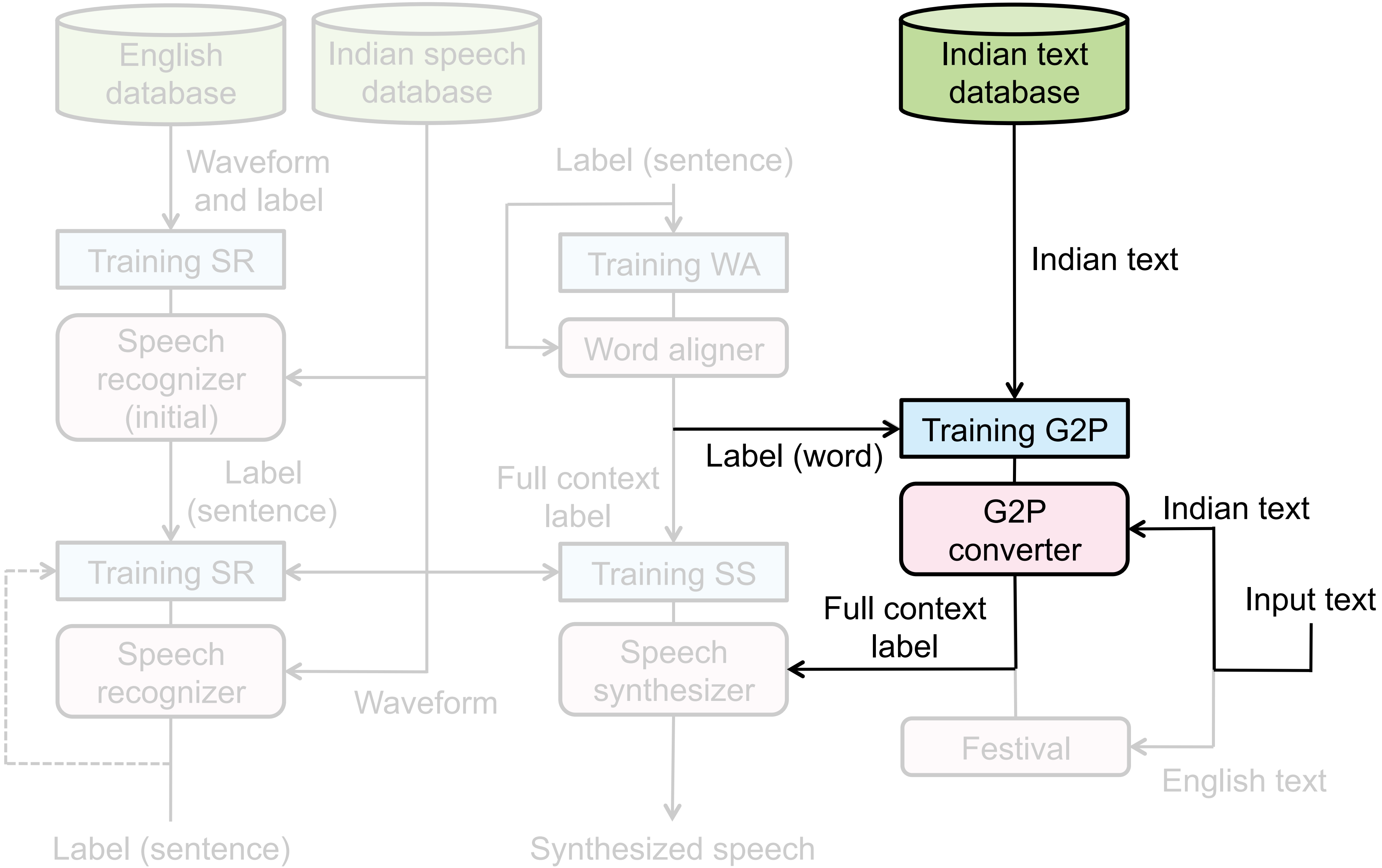
STRAIGHT [Kawahara, et al.]

- High quality speech vocoding method
- F0, spectrum, and aperiodicity measures

GV [Toda, et al.]

- Intra-utterance variance of speech-parameter trajectory

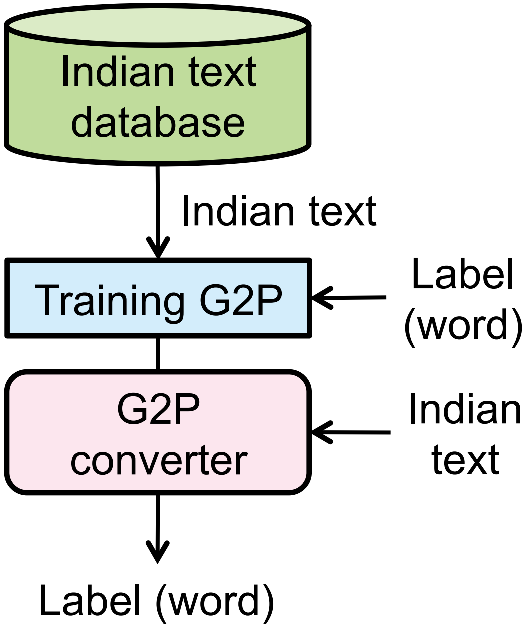
System overview



Grapheme-to-phoneme (G2P)

Joint multigram G2P converter [Bisani, el at.]

- Optimal grapheme and phoneme pair alignment is estimated
- Joint multigram models are estimated by using EM algorithm
- G2P converter is trained by using Sequitur G2P



⇒ Obtain label sequences of input text of target Indian language

Advantage of our system

Multilingual speech synthesis

- Phoneset of acoustic model is the same as the English speech recognizer
- Available text analysis results of the English
- English text analysis: Festival
- Indian language text analysis: G2P converter

Language-independent

- Can apply to languages in which sentences written with a space between words
- e.g. Indian language, Spanish, Arabic

Outline

Background

Blizzard Challenge 2014 rules

System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

Experiments

Conclusions

Speech recognition conditions

English database	WSJ0, WSJ1, and TIMIT
Indian database	Six Indian language
Window	Hamming window
Frame length	25 ms
Frame shift	10 ms
Feature vector	12-dimension MFCC + Δ + ΔΔ (39 dimension)
HMM	3-state left-to-right HMM without skip transition
Insertion penalty	−30.0
Number of iteration (target language SR)	2

Speech synthesis conditions

Sampling rate	16.0 kHz
Window	f0-adaptive Gaussian window
Feature vector	39-dimension STRAIGHT mel-cepstrum, log f0, 19 aperiodicity measure + Δ + ΔΔ (183 dimension)
HMM	5-state left-to-right MSD-HSMM without skip transition

	Assamese	Gujarati	Hindi	Rajasthani	Tamil	Telugu
Number of sentences	1427	450	875	1369	822	1470
Time	2h3m11s	2h1m33s	2h0m31s	2h13m22s	1h57m48s	3h6m32s

Evaluation conditions

Evaluation criteria	Intelligibility (WER), similarity (MOS), naturalness (MOS)
System A	Natural speech
System C	NITECH system

	Assamese	Gujarati	Hindi	Rajasthani	Tamil	Telugu
Number of listeners	115	50	109	110	109	106

- SUS: semantically unpredictable sentences
- RD: read text
- ML: multilingual sentences (Indian and English)

Word error rates (SUS)

Language System	Assamese	Gujarati	Hindi	Rajasthani	Tamil	Telugu
A	51	24	22	62	32	40
B	86	34	26	100	33	55
C	**84**	**59**	**40**	**67**	**64**	**77**
D	69	40	24	65	38	54
E	76	23	27	60	37	51
F	67	25	24	64	37	46
G	74	37	29	59	37	51
H	-	41	30	67	60	57
I	69	44	30	57	34	62
J	-	-	-	-	44	-
K	-	-	25	-	-	-

Similarity and Naturalness (RD)

Language System	Assamese		Gujarati		Hindi		Rajasthani		Tamil		Telugu	
A	3.3	4.7	2.9	4.7	4.3	4.5	4.4	4.2	4.0	4.6	4.5	4.9
B	1.8	2.1	3.0	2.6	2.4	2.0	2.6	2.3	2.0	2.3	1.7	2.0
C	**2.8**	**3.3**	**3.0**	**2.8**	**2.6**	**2.5**	**3.5**	**3.3**	**2.6**	**2.7**	**2.6**	**3.1**
D	3.2	3.5	2.7	2.8	4.0	3.6	3.6	3.7	3.0	3.2	2.5	3.5
E	2.6	2.9	3.5	3.5	3.2	3.1	3.6	3.7	2.7	2.9	2.3	3.1
F	2.9	3.4	2.8	3.4	3.4	3.2	4.0	3.9	2.7	3.4	3.3	4.0
G	3.2	3.9	3.7	3.8	3.4	3.7	3.7	3.9	3.8	3.6	3.9	4.2
H	-		3.5	2.5	2.1	2.2	3.1	3.1	3.2	2.7	1.4	1.9
I	1.8	2.1	2.8	2.7	3.1	2.2	3.3	3.2	1.8	2.6	2.9	2.3
J	-		-		-		-		3.1	2.6	-	
K	-		-		2.4	3.4	-		-		-	

Left: MOS of similarity   Right: MOS of naturalness

Similarity and Naturalness (ML)

Language System	Assamese		Gujarati		Hindi		Rajasthani		Tamil		Telugu	
A	3.8	4.9	3.7	4.7	3.7	4.3	4.3	4.3	4.0	4.6	4.7	4.9
B	1.6	1.9	2.7	3.0	1.9	2.0	2.2	2.3	2.2	2.3	1.6	2.0
C	**2.5**	**2.8**	**2.5**	**2.6**	**2.7**	**2.6**	**3.4**	**3.3**	**3.1**	**2.6**	**2.4**	**2.5**
D	2.8	2.7	2.3	2.5	3.3	2.8	3.4	3.6	3.1	3.2	3.0	3.1
E	2.3	2.2	3.5	2.9	2.5	2.6	3.4	3.7	2.6	2.8	2.6	3.1
F	-		-		1.9	2.8	3.1	3.2	-		1.9	2.3
G	-		-		-		-		-		-	
H	-		-		-		-		-		-	
I	-		-		-		-		-		-	
J	-		-		-		-		2.7	2.8	-	
K	-		-		2.4	3.0	-		-		-	

Left: MOS of similarity   Right: MOS of naturalness

Speech samples

	Assamese	Gujarati	Hindi	Rajasthani	Tamil	Telugu
RD	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊
SUS	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊
ML	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊

আ inter[illegible]usand

found [illegible] ယ

પાસેથી [illegible]டும்

- Generate multilingual speech
- Need to improve intelligibility

Outline

Background

Blizzard Challenge 2014 rules

System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

Experiments

Conclusions

Conclusions

TTS developed for Blizzard Challenge 2014

- System was built without the phoneme information and phoneset of target Indian language
- Can apply to languages in which sentences written with a space between words
- Generate multilingual speech
- Generate low intelligible speech
  - There is still room for improvement

Future work

- Improve accuracy of G2P converter
- Evaluation in other languages

References

[Hunt, et al.]: A. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proceedings of ICASSP 1996, vol. 1, pp. 373–376, 1996.

[Tokuda1, et al.]: K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proceedings of IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.

[Zen1, et al.]: H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proceedings of ICASSP 2013, pp. 7962–7966, 2013.

[Black, et al.]: A. W. Black and K. Tokuda, “The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets,” Proceedings of Interspeech 2005, pp. 77–80, 2005.

[Deligne, et al.]:S. Deligne and F. Bimbot, “Language modeling by variable length sequences : Theoretical formulation and evaluation of multi- grams,” Proceedings of ICASSP 1995, pp. 169–172, 1995.

[Zen2, et al.]: H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden semi-Markov model based speech synthesis,” Proceedings of ICSLP, pp. 1185–1180, 2004.

[Tokuda2, et al.]: K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multi-space probability distribution HMM,” IEICE Transactions on In- formation & Systems, vol. E85-D, no. 3, pp. 455–464, 2002.

[Kawahara, et al.]: H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187–207, 1999.

[Toda, et al.]: T. Toda and K. Tokuda, “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” Proceedings of Interspeech 2005, pp. 2801–2804, 2005.

[Bisani, el at.]: M. Bisani and H. Ney, “Joint-sequence models for grapheme- to-phoneme conversion,” Proceedings of Speech Communication, vol. 50, no. 5, pp. 434–451, 2008.

Thank you