Building Speech Synthesis Systems for Indian Languages

Hema A Murthy
hema@cse.iitm.ac.in

TTS Consortium, Funded by DeitY, GoI, India

03 July 2017

Outline

1. Text to Speech Synthesis
2. Languages of India
3. Data Collection
4. Segmentation
5. Syllables and group delay functions
6. DNN/CNNs and boundary correction
   - Experiments and Results
   - Architectures - Sub-utterance level
   - Experiments and Results
7. Scope for Improvement

Text to Speech Synthesis

Given an input text in a particular language, the objective of a TTS system is to produce natural and intelligible speech output.

Applications

- Readers for the visually challenged.
- Readers for small form factor smart phones.
- Enablers for the language challenged.

Objective: Build TTS systems for a number of Indian languages – with a short turnaround time for any new Indian language.

What does speech synthesis involve?

- Collection of speech data with correct transcriptions.
- The text must consist of examples of all sub-word units in various contexts.
- The speech needs to be segmented into sub-word units.
- The sub-words (or their models) are stored in a database.
- During synthesis
  - A sentence is split into subword units.
  - Waveforms (as in Unit Selection Synthesis) or models (as in Statistical Parametric Synthesis) of sub-words (based on context) are concatenated to generate the waveform.

Convergence and divergence of Indian languages

- Indian scripts: based on the ancient Brahmi script.
- Writing system corresponds to Aksharas.
- Indian scripts are syllabic (C*VC*) in nature.
- Aksharas: V, CV, CCV, CCCV.
- Aryan and Dravidian – originally less similar have become more similar.
- Basic set of phones:50; 35-38 consonants and 15-18 vowels.
- Differences in these languages mostly due to phonotactics – not dialectal variations.
- Although some languages can even be characterised by a common syllable set, the prosody – duration, prominence associated with each syllable in a word can be significantly different.

Data Collection

- 13 Indian languages
- Text for optimal text selection: trisyllabic words at best, $\approx$ 60-70 thousand unique words.
- Source: Online newspapers, blogs, children's stories, avoid proper names.
- Pronunciation dictionaries – UTF-8 based – syllable-level 100,000 words.
- 5 hours of data for each language (male, female, L2 English)

Summary of data collected

Table: Statistics of the total collected text corpus for *13* Indian Languages before and after text optimization

Lng.	Raw Text				Opt. Text				
	N.sent	N.wrd	U.wrd	N.syl	N.sent	N.wrd	U.wrd	N.syl	% syl
Hin	55000	582512	54050	7100	1900	40666	9560	6193	87.23
Tam	75000	786548	68000	8340	2497	41259	22459	7184	86.13
Mar	1519950	16419046	578644	9180	5386	51881	24177	8137	88.44
Ben	50000	538124	78716	4374	7762	84206	22371	4374	100
Mal	378654	3560283	496682	13683	4232	46008	25291	6591	48.16
Tel	4643	110241	32528	1614	4643	110241	32528	1614	100
Kan	41037	365399	55479	4182	1279	16621	10321	3706	88.62
Guj	17000	202847	34876	5698	8450	104526	26573	5438	95.43
Raj	6926	67758	15455	4215	2851	33983	9978	3352	94.72
Ass	1806	32721	10487	3474	1291	27047	9476	3460	99.59
Man	2007	26028	9607	2337	963	13095	6431	2337	100
Odi	35404	737654	67678	5803	2973	29948	11151	4065	70.04
Bod	15000	162400	18436	3134	8123	144526	17152	2913	94.98

Speech Segmentation

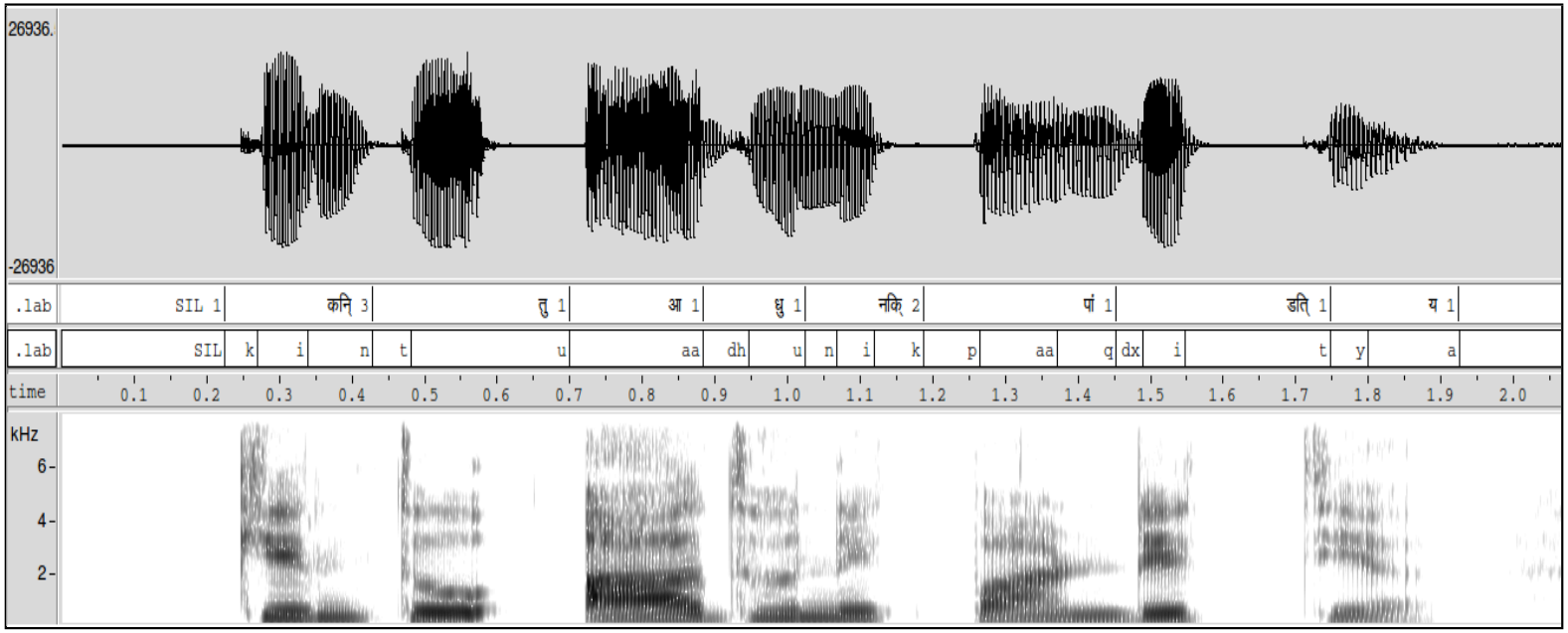


Figure: Segmentation at syllable and phone level using HMM based flat start

Essentially these segments are used in building models of phones for speech synthesis.

Question: Is it possible to improve the segmentation with small amounts of data?

Segmentation in Speech Systems

- Segmentation is important for both speech synthesis and recognition.
- Speech recognition
  - An utterance is split at the word level, which in turn is split into subwords (phone/syllable/triphone).
  - A sequence of words is obtained.
- Speech synthesis
  - A given text is split into a sequence of subword units.
  - The waveforms or models are concatenated to generate the waveform.
  - Unlike speech recognition, the consumers of synthesis are the human ears.
  - Accurate segmentation is crucial for speech synthesis.

An Aside: Parsers for Indian Languages

A uniform parser for Indian languages

- A common label set for all Indian languages
- A common set of rules
- Exception handling for each language – a set of special rules
  - taajamahala – taajmahal/taajamhal/taajamahala – depending upon the language and context.
- A lex and yacc based parser is developed for Indian languages – supports 13 Indian languages – 4 Dravidian (Tamil, Telugu, Kannada, Malayalam), 8 Aryan (Hindi, Bengali, Gujarati, Marathi, Rajasthani, Odia, Assamese, Manipuri), and 1 Sino-Tibetan (Bodo).

Mapping to Common Label Set

Hindi

Vowels				
अ	आ		इ	ई
उ	ऊ	ए	ऐ	ऐ
ओ	औ	औ		
Consonants				
क	ख	ग	घ	ङ
च	छ	ज	झ	ञ
ट	ठ	ड	ढ	ण
त	थ	द	ध	न
प	फ	ब	भ	म
Semi Vowels				
य	र	ल	ळ	व
Fricatives				
श	ष	स	ह	

Tamil

Vowels				
அ	ஆ		இ	ஈ
உ	ஊ	எ	ஏ	ஏ
ஓ	ஓ	ஔ		
Consonants				
க				ங
ச		ஜ		ஞ
ட				ண
த			ந	ன
ப				ம
Semi Vowels				
ய	ர	ல	ள	வ
Fricatives				
	ஶ	ஷ	ஸ	

Vowels				
a	aa	ax	i	ii
u	uu	e	ee	ei
o	oo	ou		
Consonants				
k	kh	g	gh	ng
ch	c	jh	j	nj
tx	txh	dx	dxh	nx
t	th	d	dh	n
p	ph	b	bh	m
Semi Vowels				
y	r	l	lx	w
Fricatives				
sh	sx	s	h	

Bengali

Vowels				
	আ	ঐ	ঈ	ঈ
উ	ঊ		ঐ	ঐ
	ঔ	ঔ		
Consonants				
ক	খ	গ	ঘ	ঙ
চ	ছ	জ	ঝ	ঞ
ট	ঠ	ড	ঢ	ণ
ত	থ	দ	ধ	ন
প	ফ	ব	ভ	ম
Semi Vowels				
য	র	ল		
Fricatives				
শ	ষ	স	হ	

Malayalam

Vowels				
അ	ആ		ഇ	ഈ
ഉ	ഊ	ഋ	ഌ	ൠ
ഒ	ഓ	ഔ		
Consonants				
ക	ഖ	ഗ	ഘ	ങ
ച	ഛ	ജ	ഝ	ഞ
ട	ഠ	ഡ	ഢ	ണ
ത	ഥ	ദ	ധ	ന
പ	ഫ	ബ	ഭ	മ
Semi Vowels				
യ	ര	ല	ള	വ
Fricatives				
ശ	ഷ	സ	ഹ	

Proposed rules[1]

Hindi words : अकबर असफल follows v-cv-cv-cv structure

But अकबर (a-k-a-b-a-r-a) parsed as a-**k-b-a**-r (vc-cvc)

and असफल (a-s-a-f-a-l-a) parsed as a-**s-a-f-a**-l (v-cv-cvc)

Hindi words : ताजमहल पागलपन follows cv-cv-cv-cv-cv structure

But ताजमहल (t-aa-j-a-m-a-h-a-l-a) parsed to t-aa-**j-m-a-h-a**-l (cvc-cv-cvc)

and पागलपन (p-aa-g-a-l-a-p-a-n-a) parsed to p-aa-**g-a-l-p-a**-n (cv-cvc-cvc)

[1] Arun Baby N L Nishanthi, Anju L Thomas and Hema A. Murthy, "A Unified parser for developing Indian language text to speech synthesizers", in *International Conference on Text, Speech and Dialogue (TSD)*, Sept 2016, pp. 514-521.

Parsing of agglutinative words

- Agglutination is the process of combining words that are formed by **stringing** together morphemes
- Unified parser handles the agglutinative words that are common in Dravidian languages since it employs a **rule-based approach**

Tamil

வந்துகொண்டிருக்கிறான் → வந்து கொண்டு இருக்கிறான்

w-a-nd-d-u-k-o-nx-dx-i-r-u-k-k-i-rx-aa-n → w-a-nd-d-u k-o-nx-dx-u i-r-u-k-k-i-rx-aa-n

Malayalam

വന്നുകൊണ്ടിരിക്കുന്നു → വന്നു കൊണ്ട് ഇരിക്കുന്നു

w-a-n-n-u-k-o-nx-tx-i-r-i-k-k-u-n-n-u → w-a-n-n-u k-o-nx-tx i-r-i-k-k-u-n-n-u

Building speech synthesis systems for Indian languages

Bootstrap Segmentation

- Small amount of data labeled manually at the phone level.
- Phone models are built.
- Forced Viterbi alignment is performed on the rest of the data using the models built.
- Models are rebuilt using the forced aligned data.

Issues: Inconsistencies

- Perceiving phones based on listening and spectrogram is difficult in isolation.
- Inconsistency across annotators.

Flatstart Segmentation

- HMM models initialised such that state means and variances are equal to the global mean and variance.
- Embedded training[2] is performed to build models:
  - Uses transcription to obtain composite HMM for each utterance by concatenating phone HMMs.
  - Embedded Baum-Welch re-estimation.
- Using these models, forced Viterbi alignment is performed to obtain segmentation at phone level.

[2] *HTK: Speech recognition toolkit*, http://htk.eng.cam.ac.uk/.

Drawbacks of HMM Based Segmentation

- A fundamental drawback of this approach is that boundaries are not explicitly modeled [3]
- HMMs do not use proximity to boundary positions as a criterion for optimality during training [4]

[3] Yuan et al. "Automatic phonetic segmentation using boundary models," INTERSPEECH 2013.

[4] Sethy et al., "Refined speech segmentation for concatenative speech synthesis," ICSLP, 2002.

Example of segmentation at syllable-level is FS HMM (at phone-level)

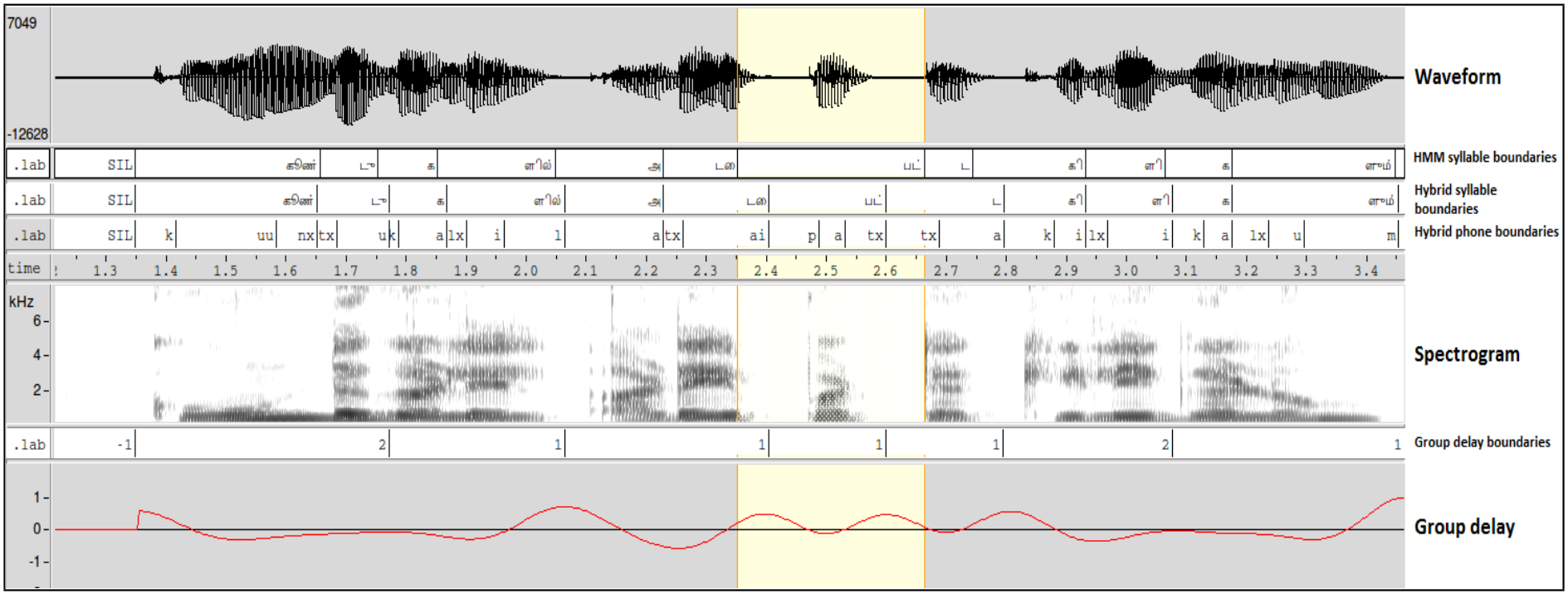


Figure: Example

Syllable /p a tx/

Observe the boundary for the stop consonants at either end is wrong as evidenced by the spectrogram.

Boundary Correction Techniques

- Support Vector Machine (SVM) classifiers are used to locate boundaries in Hung-Yi Lo et al., "Phonetic boundary refinement using support vector machines," ICASSP, 2007. A special one-state HMM is used for detecting phoneme boundaries in [5]
- In [6] a multi layer perceptron is used to refine phone boundaries.

[5] Yuan, Jiahong and Ryant, Neville and Liberman, Mark and Stolcke, Andreas and Mitra, Vikramjit and Wang, Wen,"Automatic phonetic segmentation using boundary models.," pp.2306-2310, INTERSPEECH 2013.

[6] Ki-Seung Lee, "MLP-based phone boundary refining for a TTS database," IEEE Trans, ASLP, Vol.14, No.3, 2006

Drawbacks

- Supervised.
- Manually marked accurate boundaries are required for training.

Alternative

- Can we exploit time domain and spectral cues in the signal to rectify boundaries?

Energy

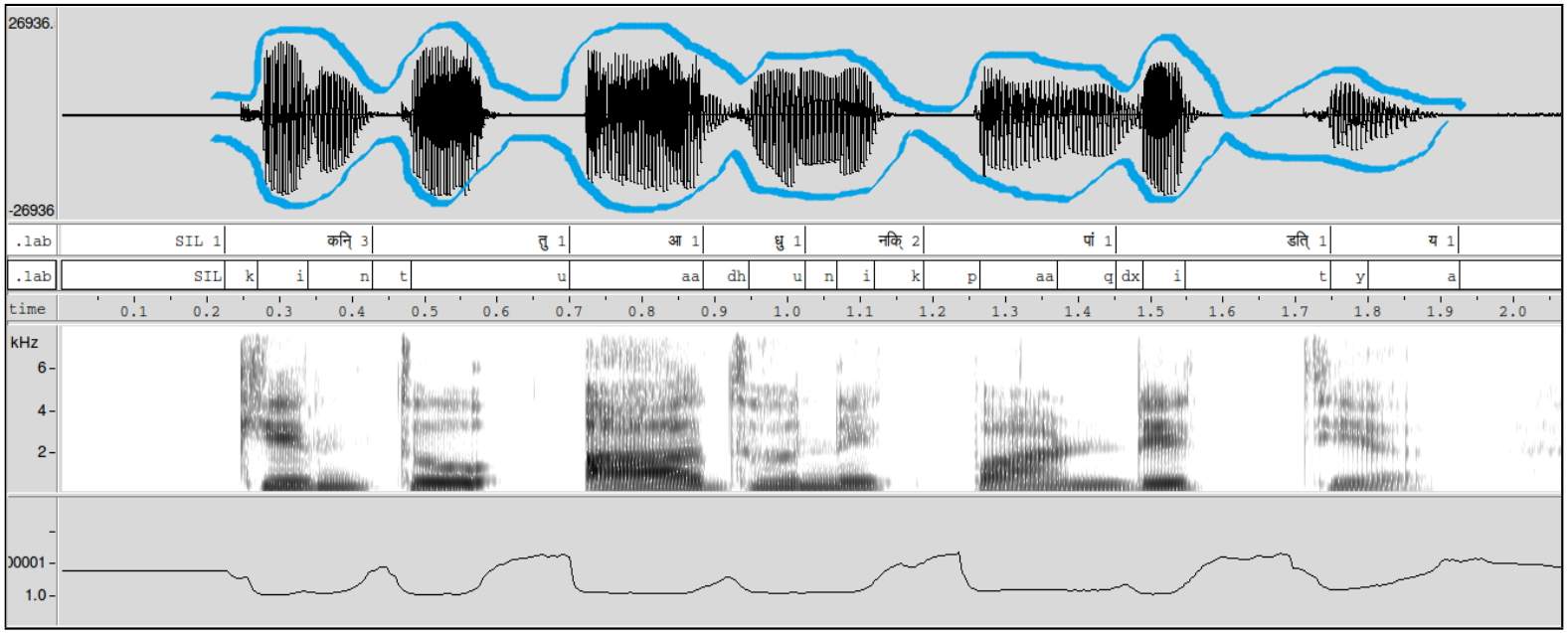


Figure: Inverse of energy as a cue

Spectral Change

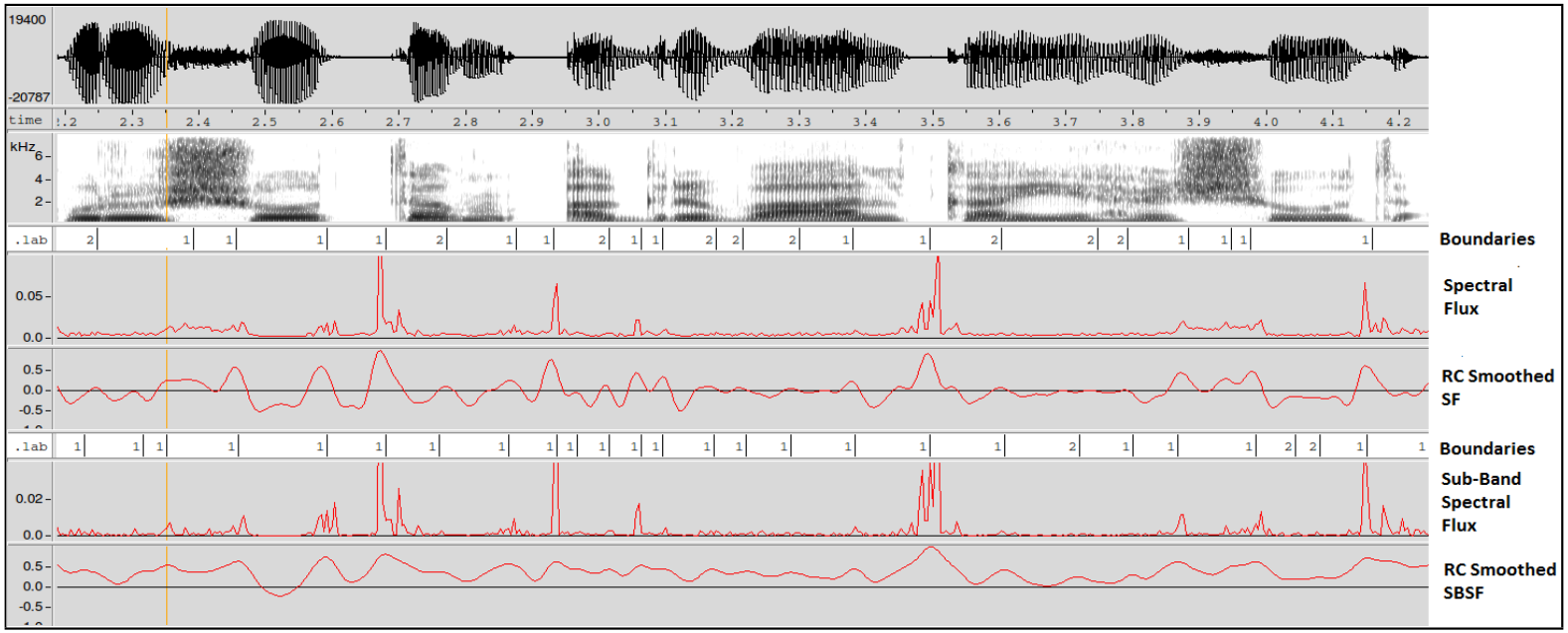
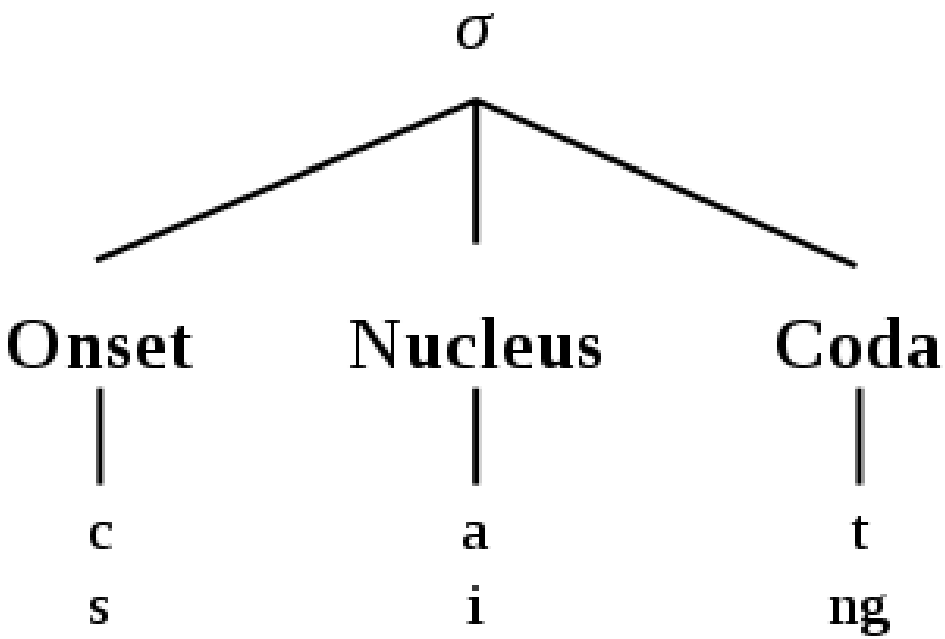


Figure: Spectral change as a cue

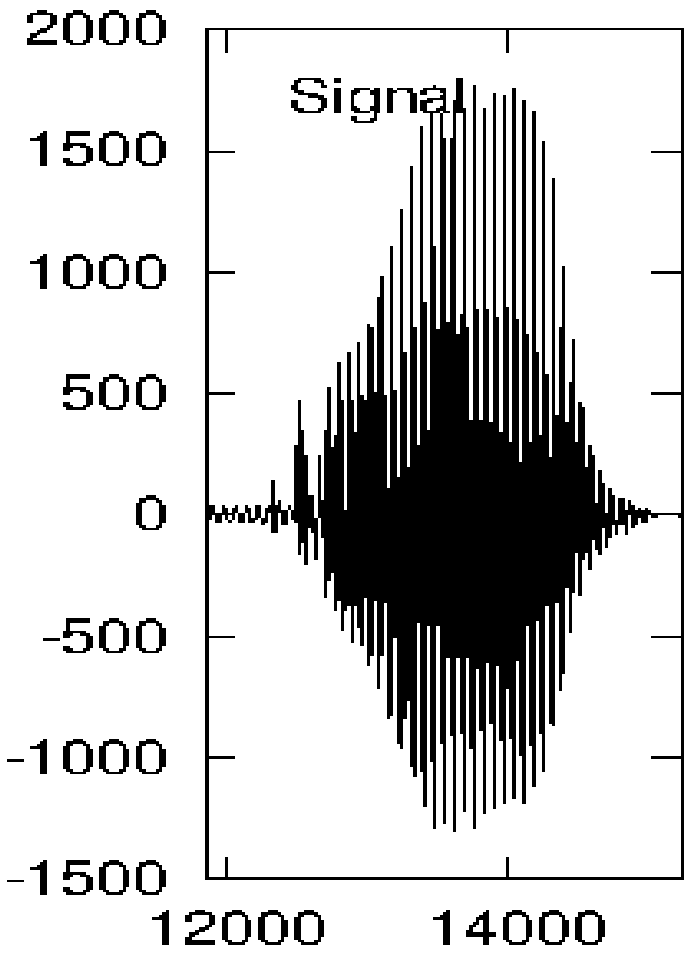
Property of a syllable

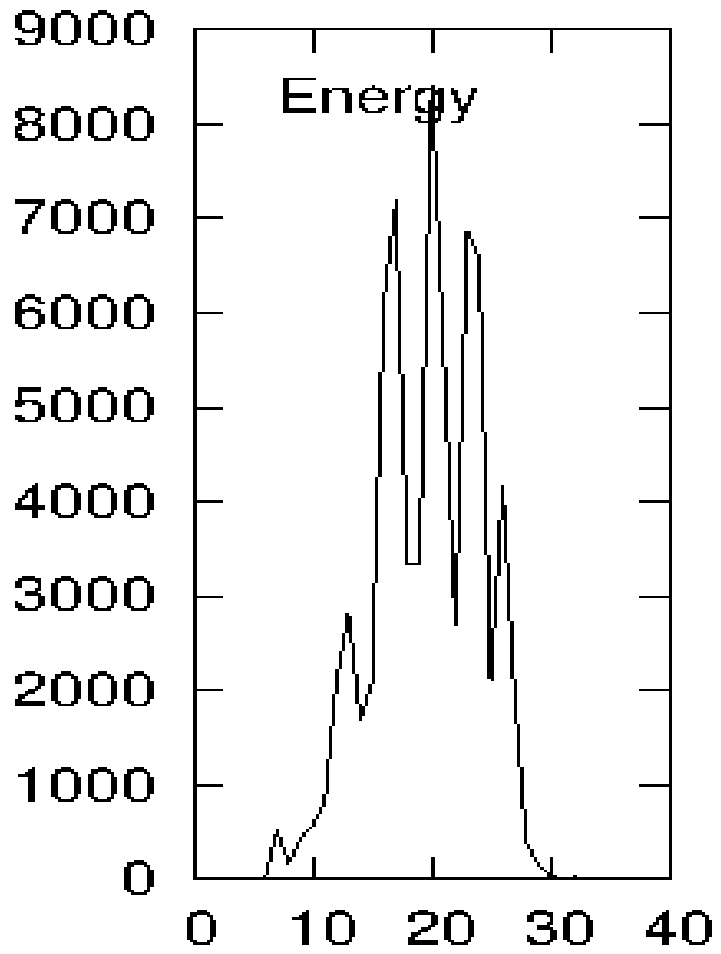
Three components: Onset, Rime (nucleus) and Coda



Rime is a sonorant, characterised by a vowel. Syllable definition: (C*VC*)

Signal Characteristics of Syllable





Energy high in the middle and tapers off towards the end.

Even English has syllables of CV type (70%) – [7]

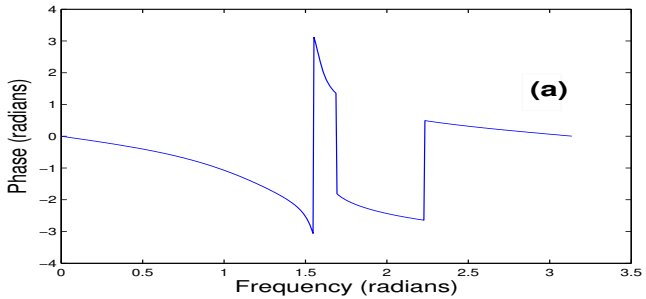
[7] Steven Greenberg, Speech Communication

Group delay functions[8]

- Consider a discrete time domain signal $x[n]$ and its Fourier transform $X(\omega) = |X(\omega)|e^{j\phi(\omega)}$

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{1}$$

- Computation of $\tau(\omega)$ directly from phase not trivial



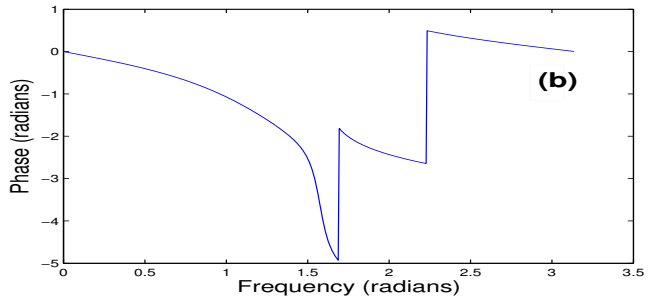


Figure: a) Wrapped, and b) Unwrapped phase response of an Elliptic low pass filter

- Alternate form of computation:

$$\tau(\omega) = \frac{X_R(\omega)\,Y_R(\omega) + X_I(\omega)\,Y_I(\omega)}{|X(\omega)|^2} \tag{2}$$

X_R,X_I : Real, Imaginary parts of $X(\omega) = FFT(x[n])$

Y_R, Y_I : Real, Imaginary parts of $Y(\omega) = FFT(nx[n])$

[8] Hema A Murthy and B Yegnanarayana, "Group delay functions and its application to speech technology," *Sadhana, Indian Academy of Sciences*, Vol 36, No. 5, pp.745-782, November, 2011.

Resolving power of group delay functions

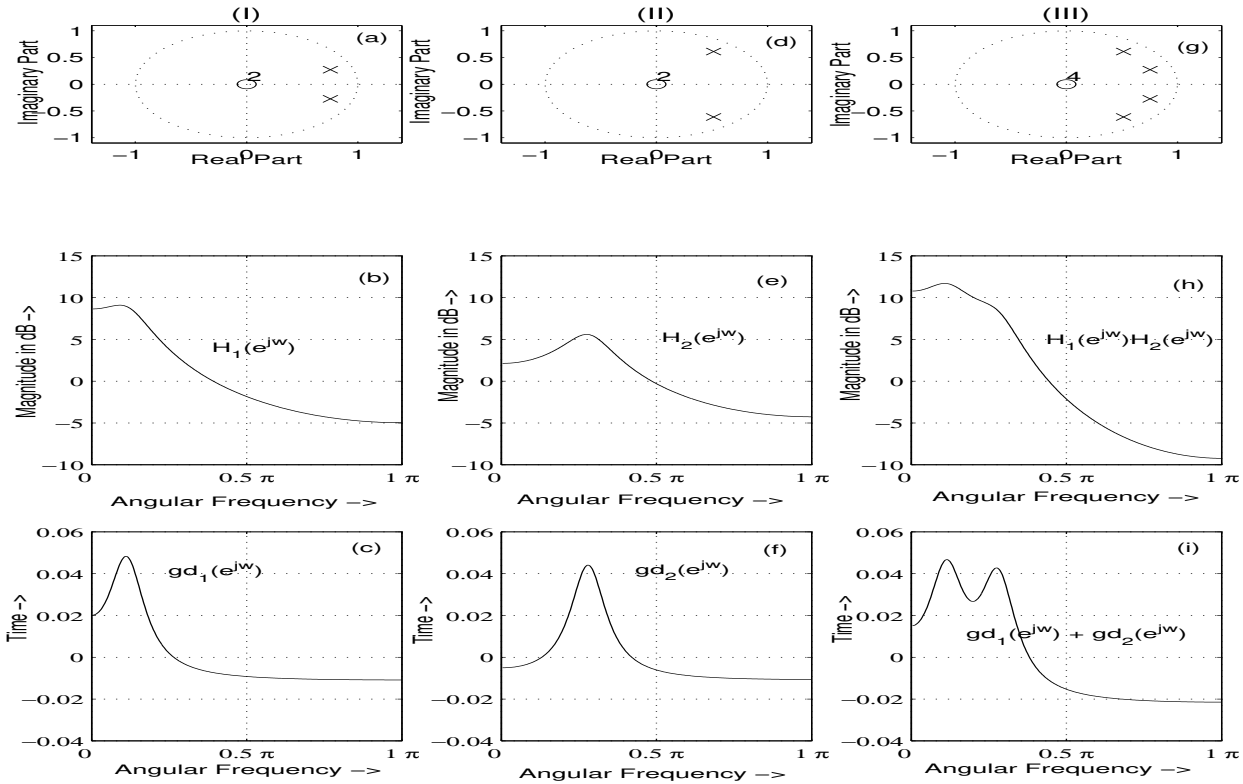


Figure: Resolving power of the group delay spectrum: z-plane, magnitude spectrum and group delay spectrum I) a pole inside the unit circle at $(0.8, \pi/8)$, II) a pole inside the unit circle at $(0.8, \pi/4)$ and III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle.

Both poles inside the unit circle $\Longrightarrow$ minimum phase system.

Mixed phase system behaviour

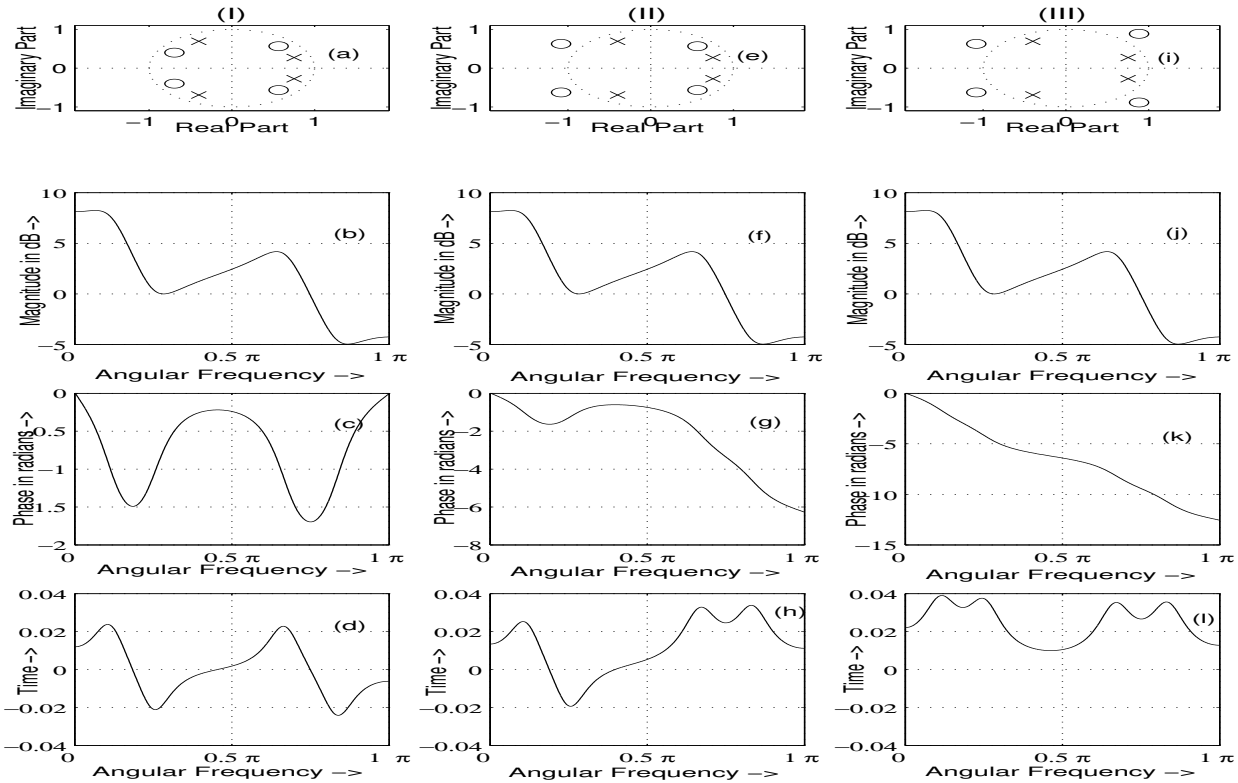


Figure: Group delay property of different types of signals: minimum and nonminimum phase signals

Clearly, system is best behaved for minimum phase system.

Revisiting the source system model for speech

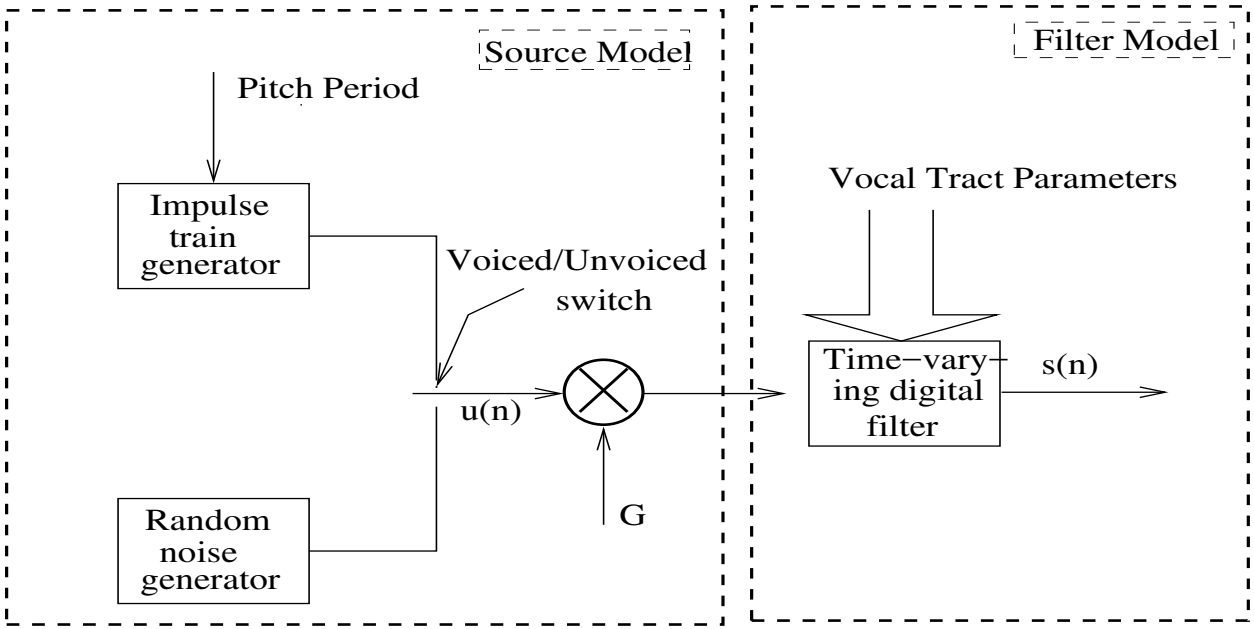


Figure: *A source system model for speech production*

- System: stable and causal $\Longrightarrow$ poles inside unit circle
- System: zeroes may lie inside or outside – nasals
- System: No zeroes on unit circle – even zeroes have finite bandwidth

Feature extraction from phase: Zeroes on the unit circle

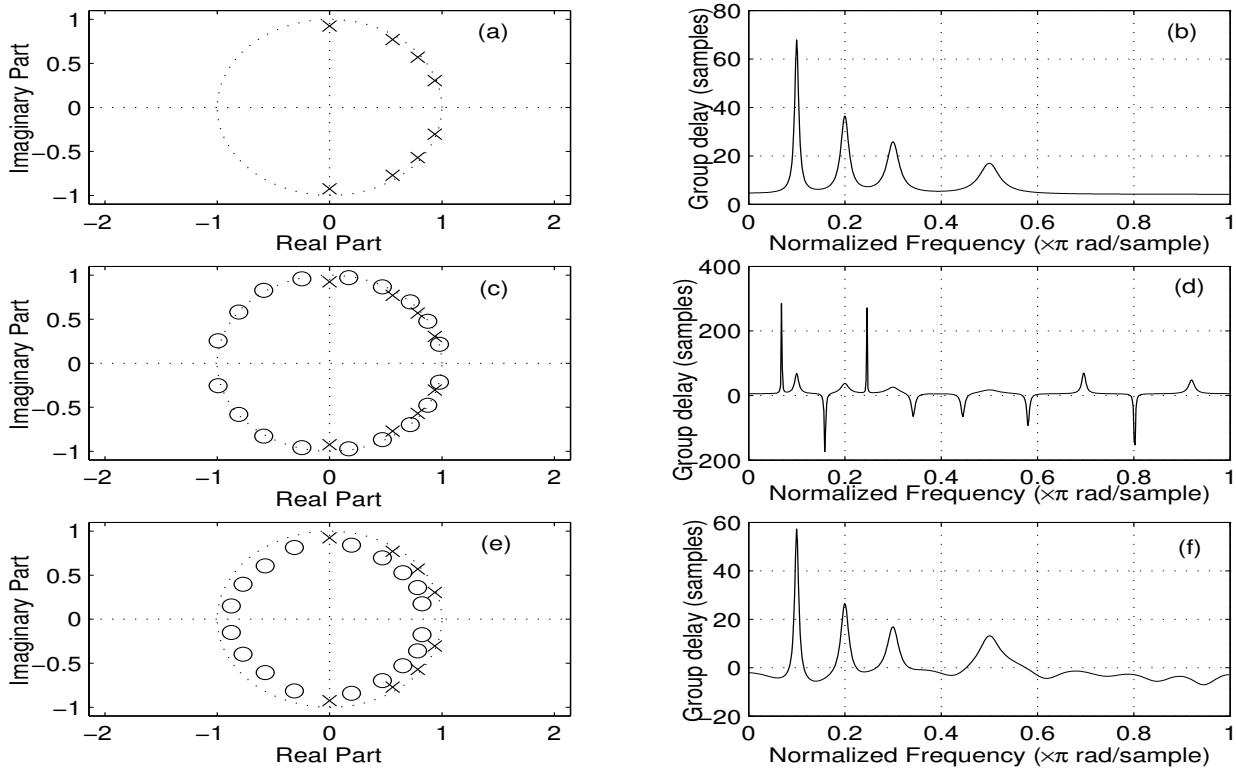


Figure: Significance of proximity of zeros to the unit circle

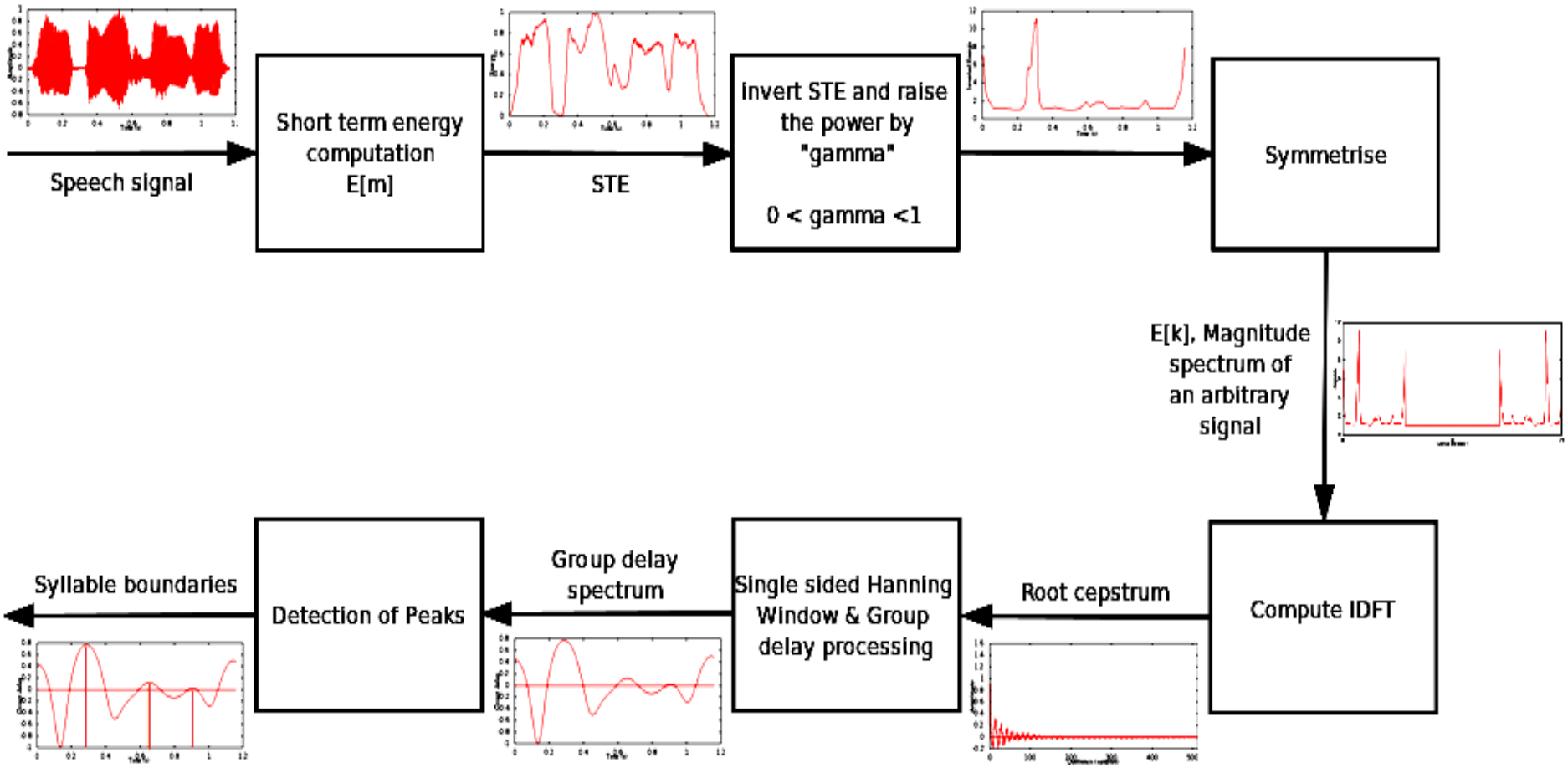
Group delay processing

- Linear prediction based group delay spectra.
- Root cepstrum based group delay spectra.
- Modified group delay spectra.

Short-term Energy as an arbitrary magnitude spectrum

- Energy is a positive function
- Symmetrise energy
- Arbitrary magnitude function
- Minimum phase group delay function
- Valleys correspond to syllable boundaries approximately

Segmentation of speech using group delay functions



Enforcing syllable boundaries during Embedded Re-estimation[10]

- Baum-Welch embedded re-estimation[9] is performed at the syllable level to build phone HMM models.
- Forced alignment also performed within the syllable

[9] S. Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2002.

[10] S Aswin Shanmugam and Hema A Murthy, *Group Delay Based Phone Segmentation for HTS*, in Proc. of Twentieth National Conference on Communications (NCC 2014)

Example

Syllable Label			Phone Label		
Beg	**End**	**Unit**	**Beg**	**End**	**Unit**
0.000	**0.234**	SIL	**0.000**	**0.234**	SIL
0.234	**0.363**	/i/	**0.234**	**0.363**	i
			0.363	0.398	s
0.363	**0.516**	/see/	0.398	**0.516**	ee
			0.516	0.576	k
0.516	**0.647**	/ka/	0.576	**0.647**	a
0.647	**0.728**	/ii/	**0.647**	**0.728**	ii
			0.728	0.807	b
			0.807	0.982	aa
0.728	**1.113**	/baar/	0.982	**1.113**	r
1.113	**1.228**	SIL	**1.113**	**1.228**	SIL

Table: Phone labels from syllable labels

Segmentation Correction

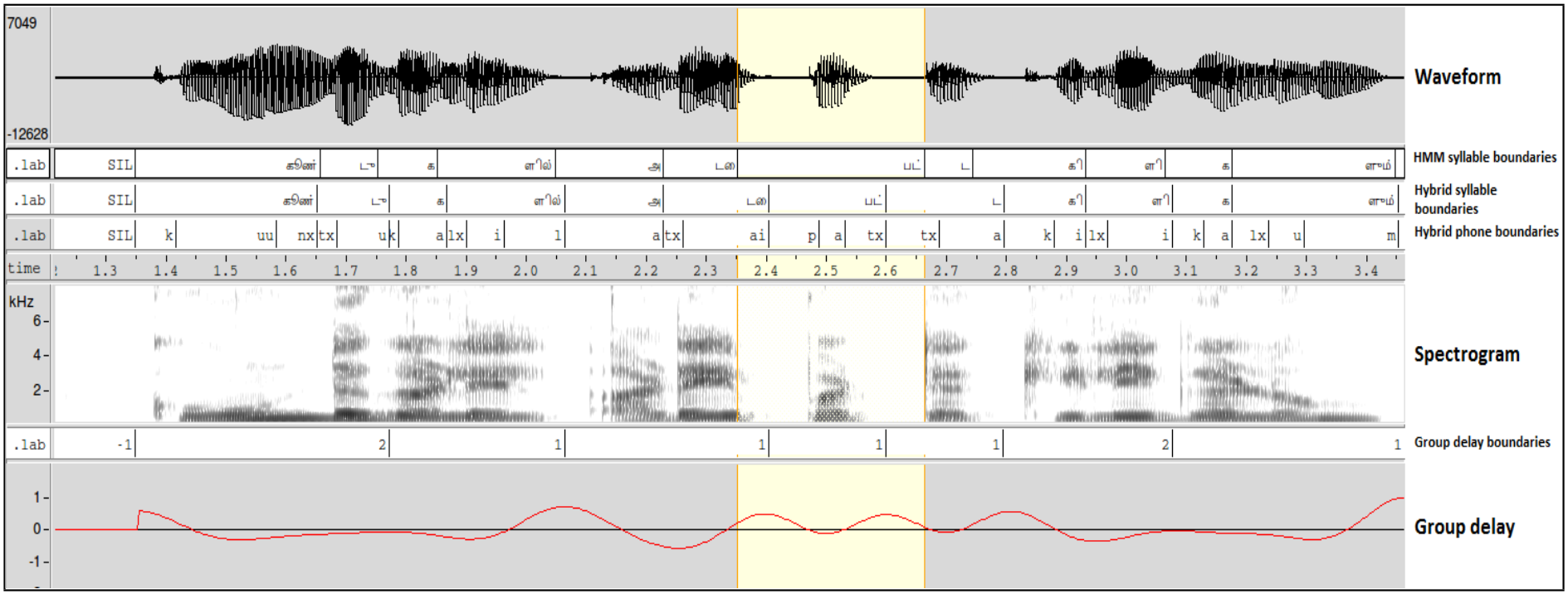


Figure: Example – without boundary correction

Syllable /p a tx/

Segmentation Correction

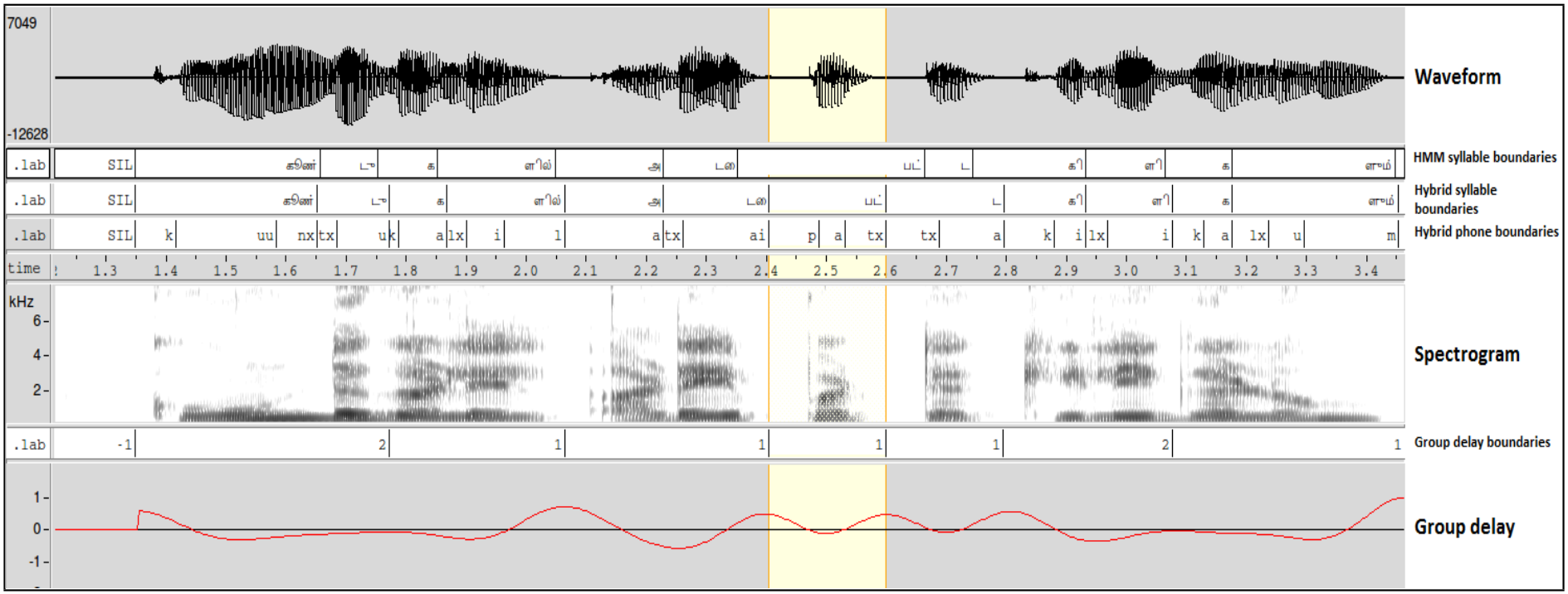


Figure: Example – with boundary correction

Syllable /p a tx/

Spectral Flux based cues

- Spectral Flux is the Euclidean distance between the normalised power spectrum of one frame and the normalised power spectrum of the previous frame

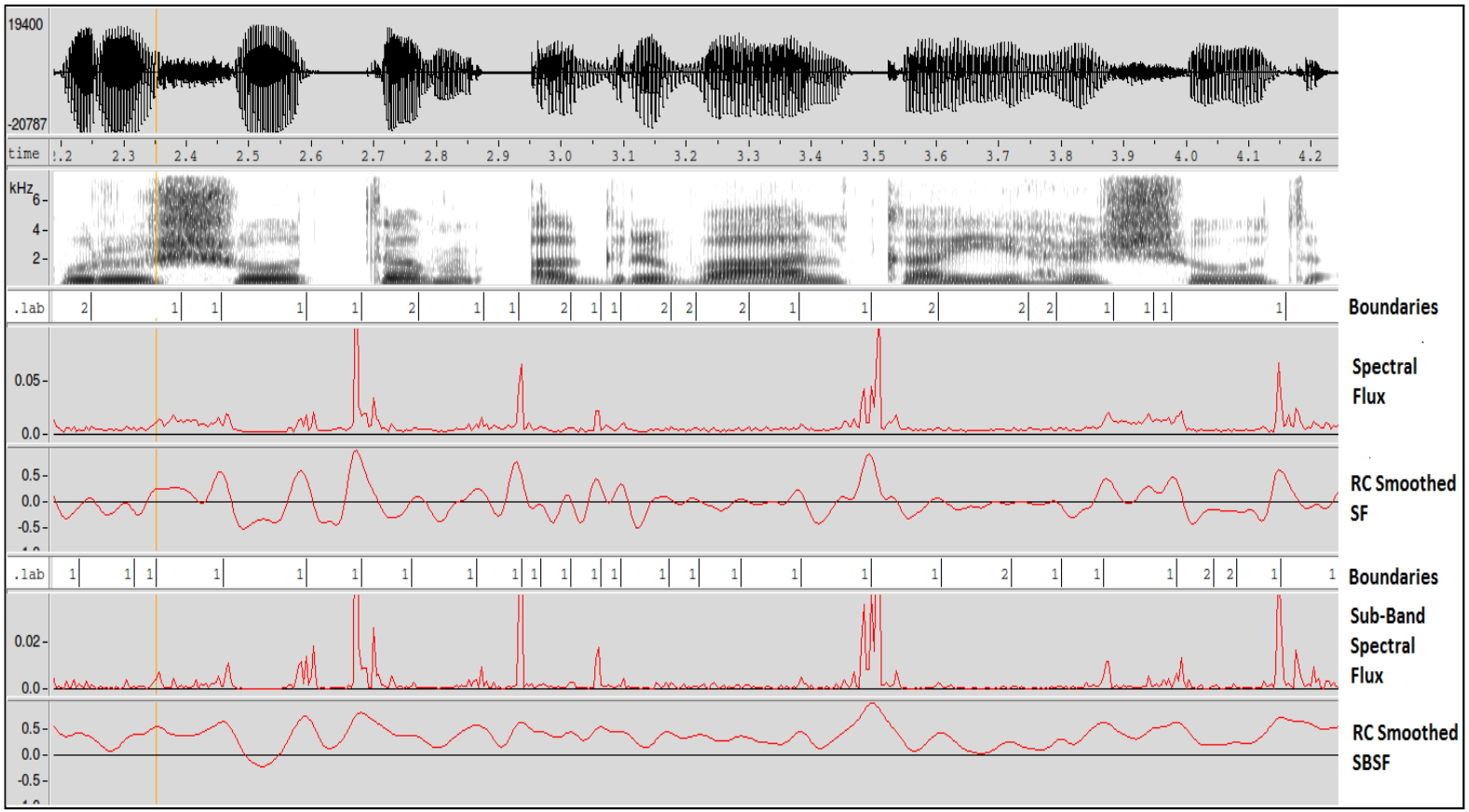


Figure: Sibilant Fricatives and Affricates Boundary Detection

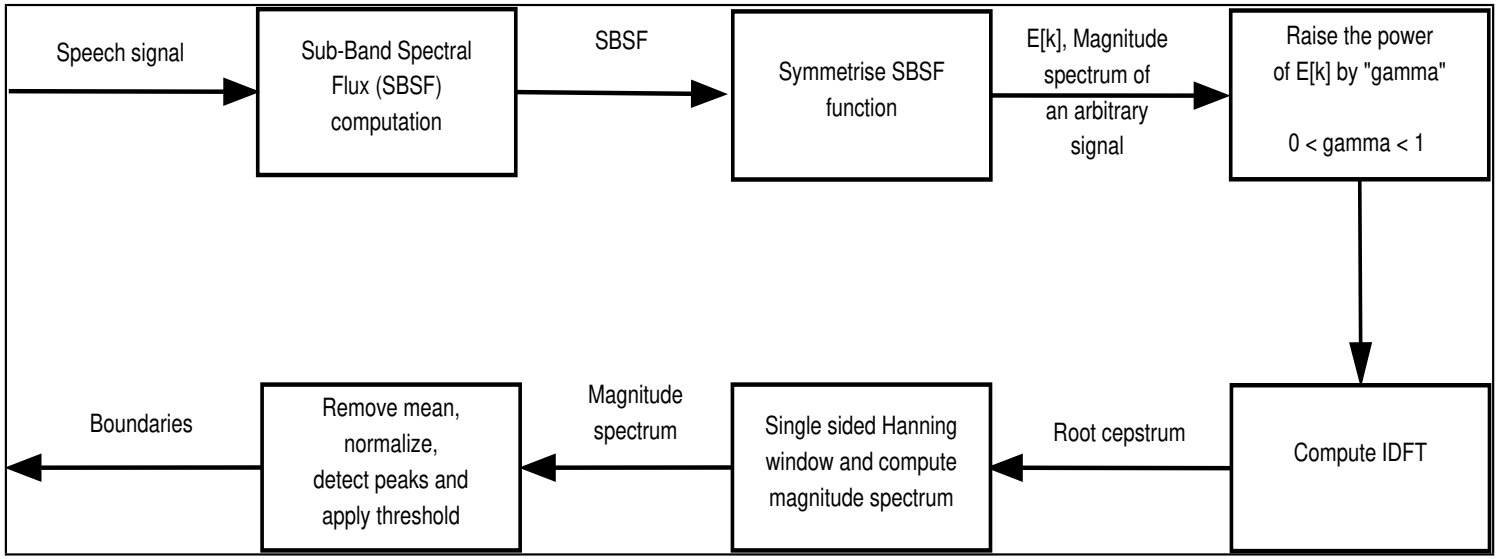


Figure: Algorithm

Hybrid Segmentation

- Uses both Short-Time Energy (STE) and Sub-Band Spectral Flux (SBSF).
- While syllable level reestimation add context to phones (eg. “d uu k” becomes “beg-d uu k_end”).
- Correction rules are different.
- Uses threshold

Boundary Correction Rules

- *syllable1* boundary *syllable2*
- $P^* e_p$ boundary $b_p P^*$
- If b_p == (Unvoiced Stop Consonant), use STE with $threshold_1$ for correction.
- If e_p == (Unvoiced Stop Consonant), use STE with $threshold_2$ for correction.
- b_p XOR e_p == (Fricative, Affricate), use SBSF with $threshold_3$ for correction.
- b_p == (Unvoiced Stop Consonant) and e_p == (Nasal), use SBSF with $threshold_3$ for correction.

Segmentation Correction

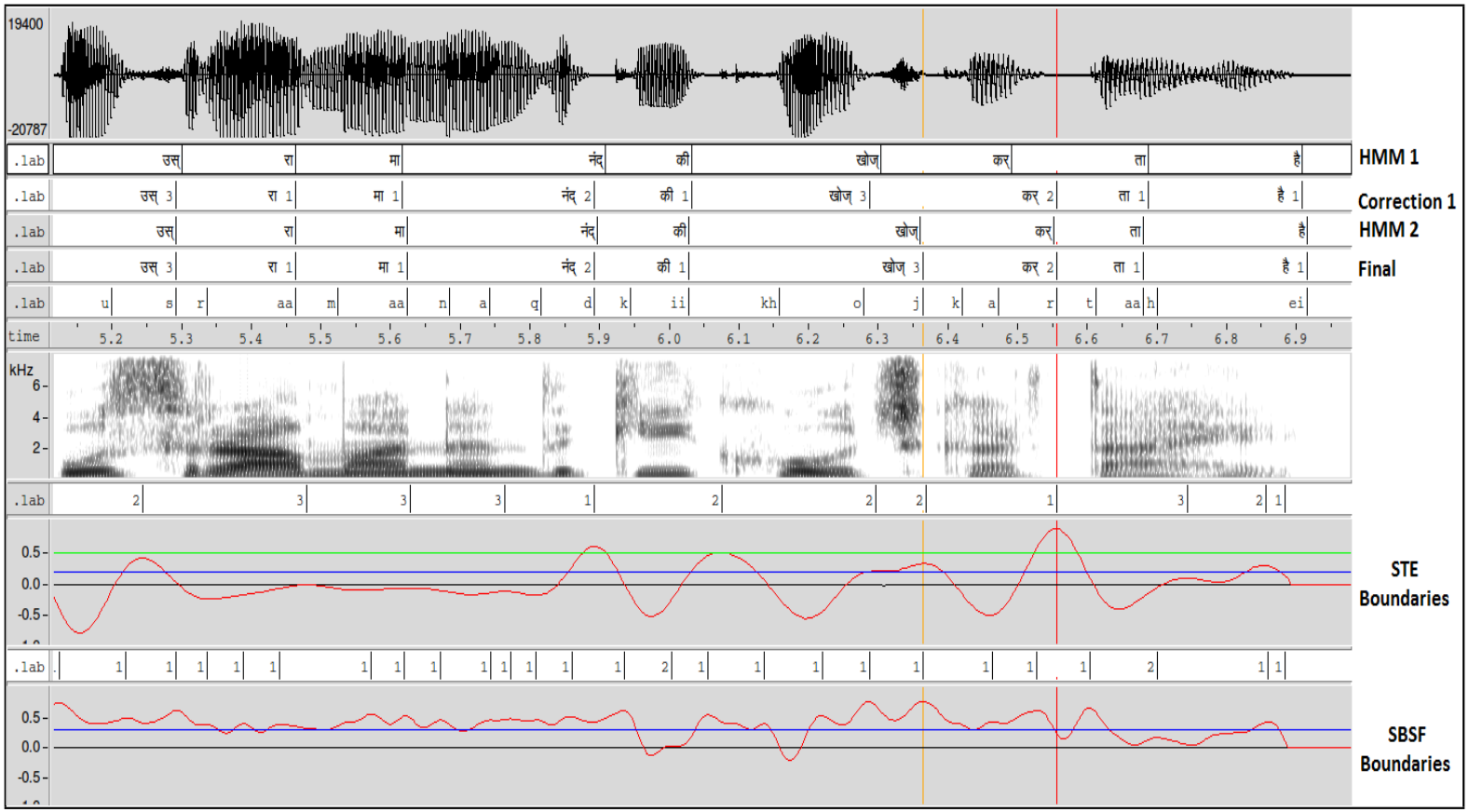
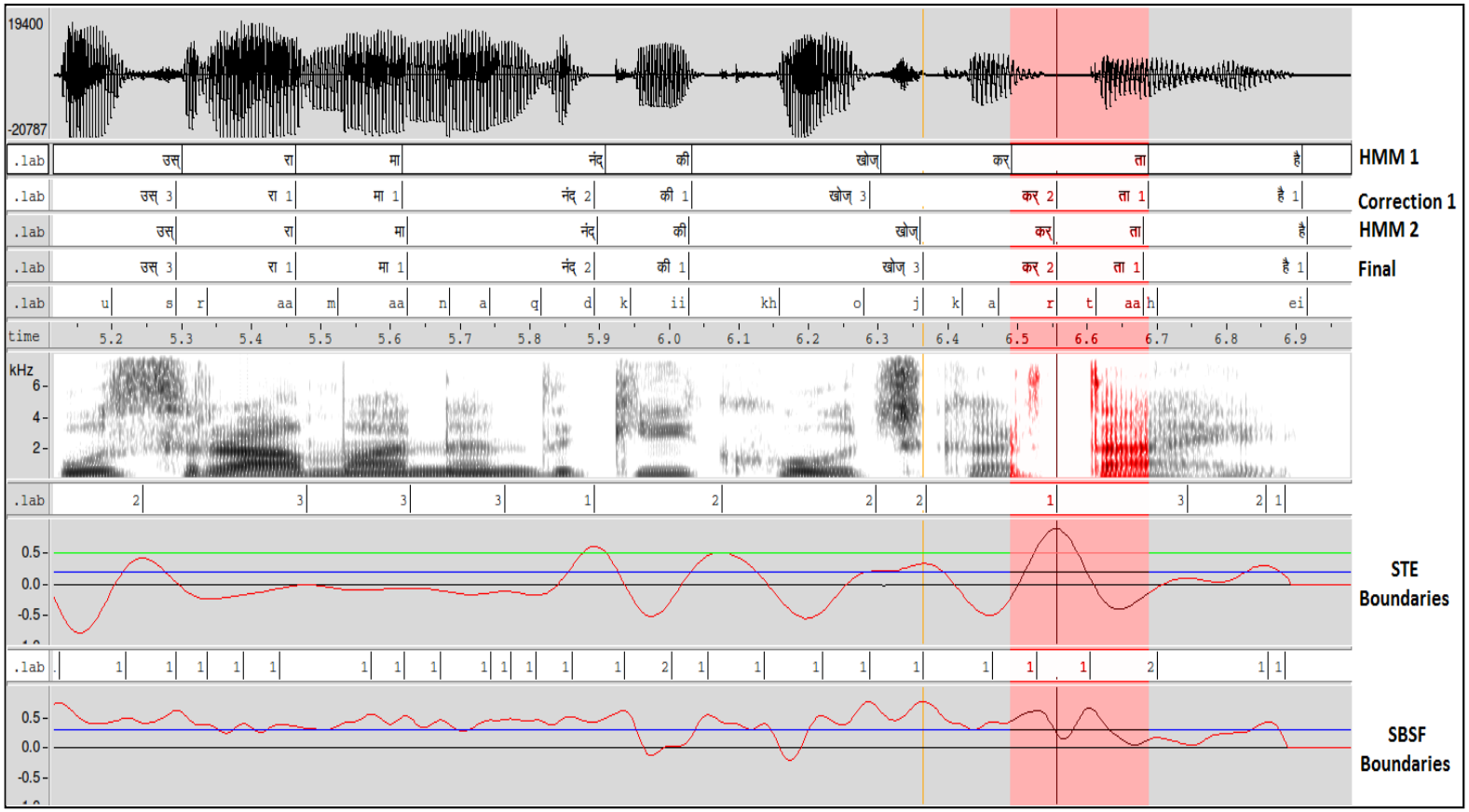


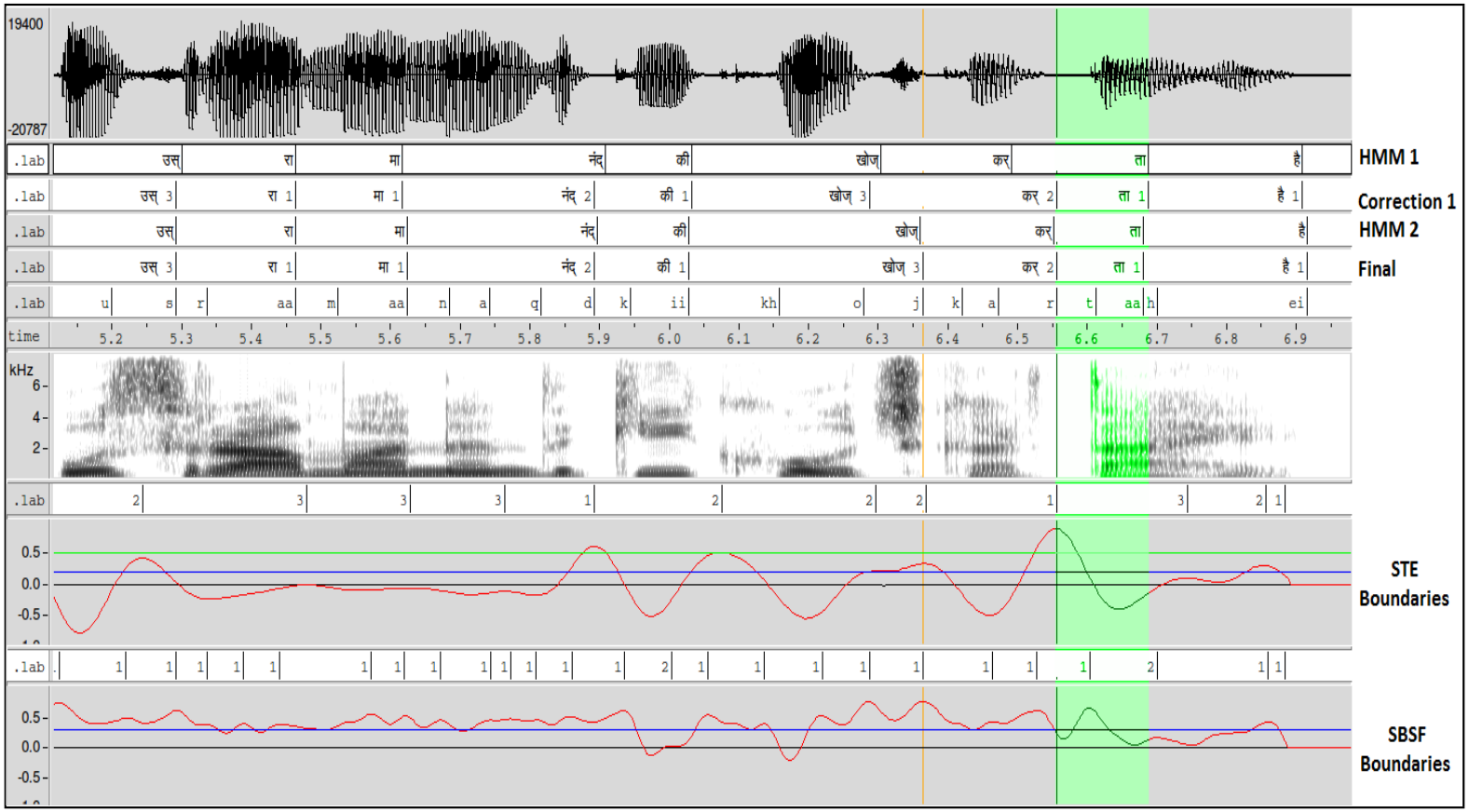
Figure: *Experiments with Hindi*

Segmentation Correction



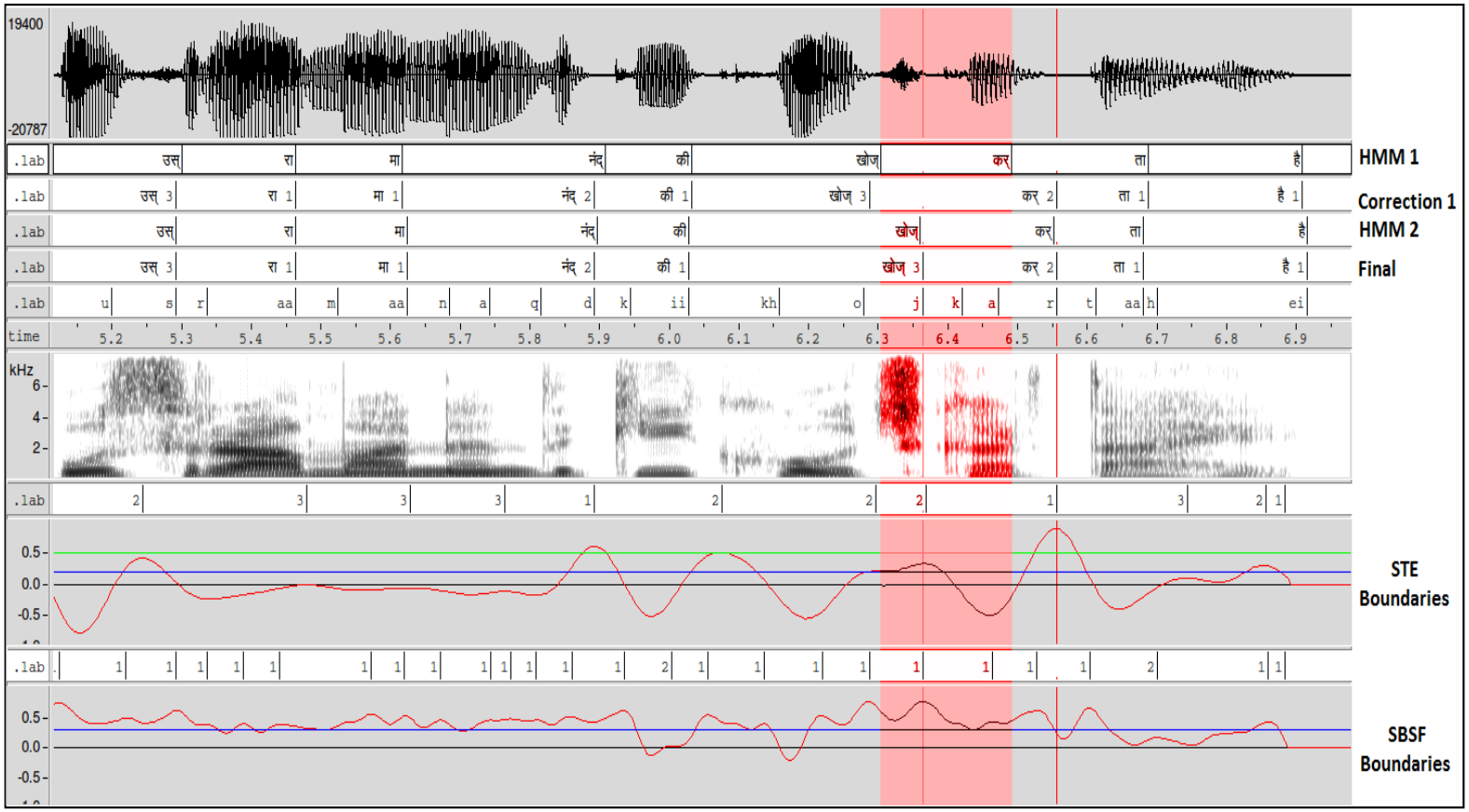
Syllable /t aa/

Segmentation Correction



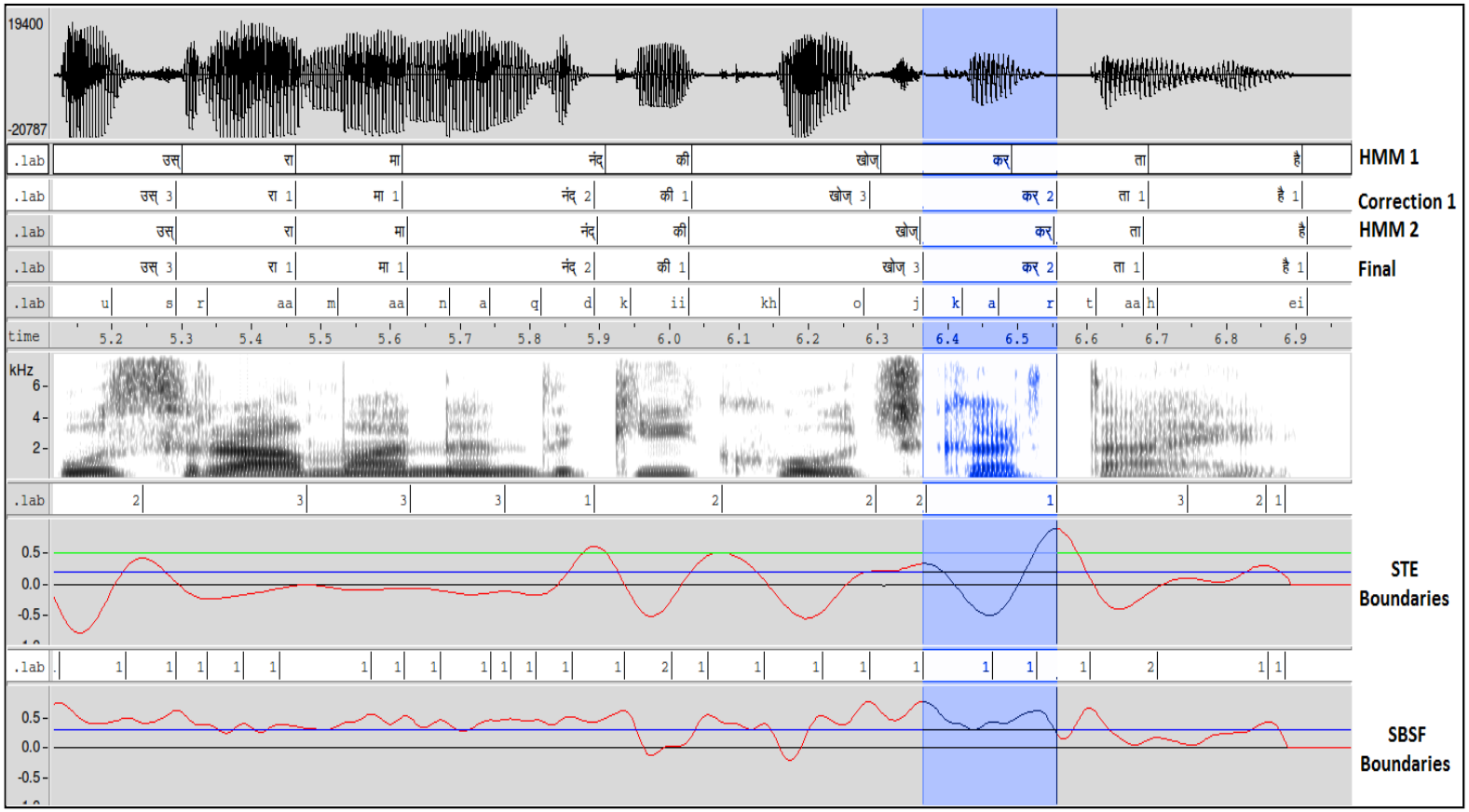
Syllable /t aa/

Segmentation Correction



Syllable /k a r/

Segmentation Correction



Syllable /k a r/

Segmentation Accuracy

Method	Average log probability per frame						
	Nasals	**Fricatives**	**Affricates**	**Semi-Vowels**	**Vowels**	**Stops Consonants**	**Overall**
Hybrid	-75.06	-77.71	-80.70	-76.90	-69.31	-83.16	-74.85
HMM-FS	-74.91	-78.50	-81.00	-77.71	-69.82	-84.00	-74.99

Table: Average log probability per frame for Hindi (Female data)

Observe that vowel likelihoods have not changed significantly, while consonants, affricates change quite significantly.

Segmentation Accuracy

Method	Average log probability per frame						
	Nasals	**Fricatives**	**Affricates**	**Semi-Vowels**	**Vowels**	**Stops Consonants**	**Overall**
Hybrid	-69.02	-77.74	-79.46	-74.16	-64.49	-82.05	-70.46
HMM-FS	-69.80	-78.24	-80.22	-74.60	-65.07	-82.82	-71.29

Table: Average log probability per frame for Hindi (Male data)

Segmentation Accuracy

Method	Average log probability per frame						
	Nasals	**Fricatives**	**Affricates**	**Semi-Vowels**	**Vowels**	**Stops Consonants**	**Overall**
Hybrid	-69.67	-79.05	-79.98	-73.99	-67.23	-83.01	-72.99
HMM-FS	-70.86	-78.48	-80.40	-74.89	-68.01	-84.41	-73.93

Table: Average log probability per frame for Tamil

Boundary correction and DNN/CNNs

Proposed Approach

Combine neural networks and signal processing

Proposed Methods

- Iterative boundary correction at utterance level

DNN configuration

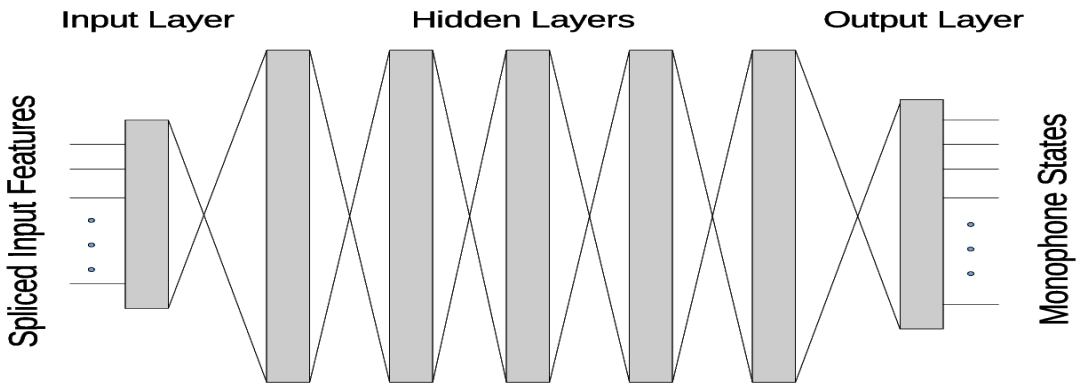


Figure: DNN configuration

- 40 dimensional fbank features over 11 frames
- RBM training to initialize the DNN weights
- Stochastic gradient descent and back propagation
- mini-batch size of 256 is used

CNN configuration

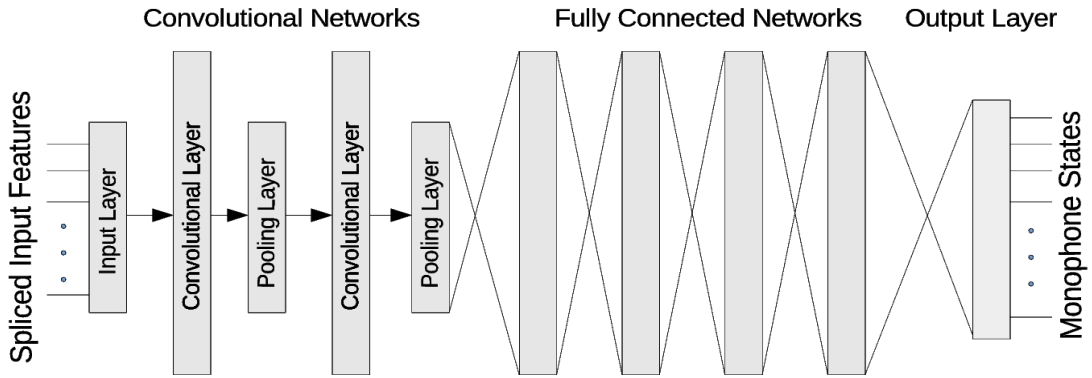


Figure: CNN configuration

- 40 dimensional fbank features with 3 pitch coefficients with a 11 frame context
- Convolutional window of dimension 8 and pooling window of size 3 with no-overlap.
- Feature map size of 256 and 128 used in first and second layers respectively

Architectures used

- HMM-GMM (Hybrid Segmentation) [11]
- HMM-DNN
- HMM-DNN with boundary correction
- HMM-CNN
- HMM-CNN with boundary correction

[11] S Aswin Shanmugam and Hema Murthy hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation. In INTERSPEECH, 2014, pp. 1648–1651

Segmentation Correction

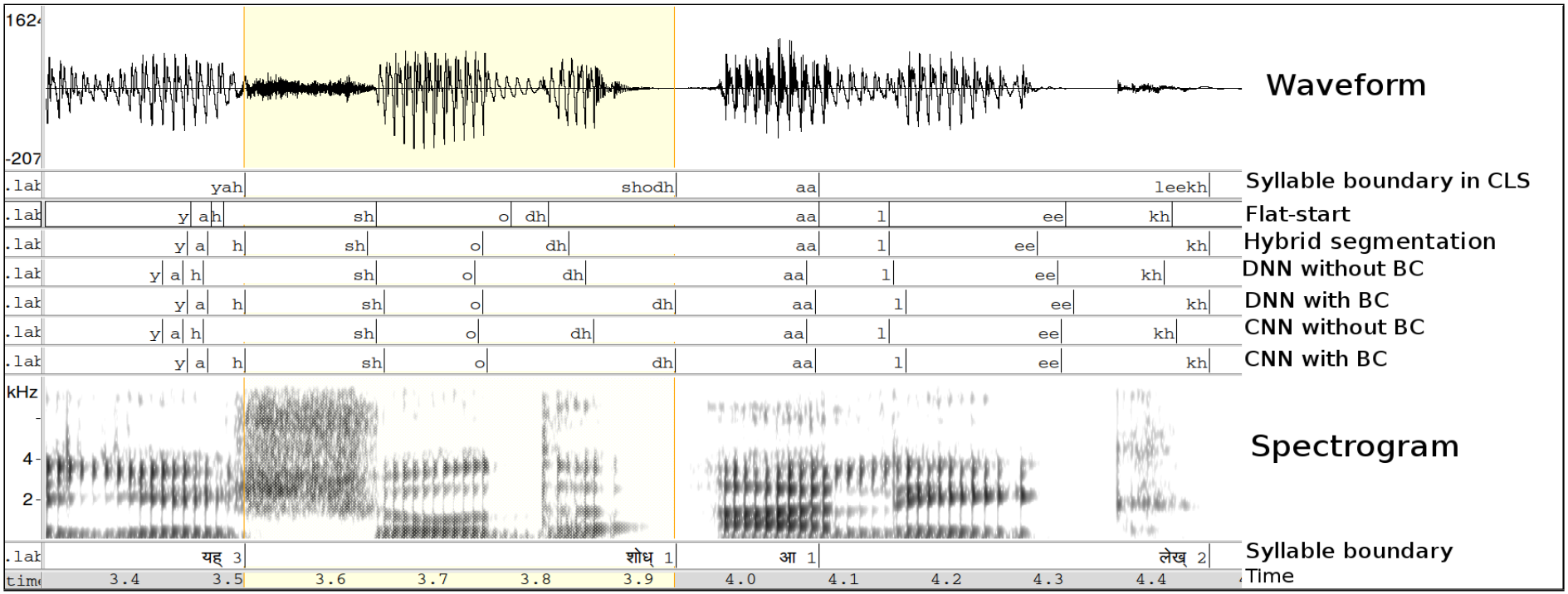


Figure: Example – manually marked

Syllable /sh o dh/

Segmentation Correction

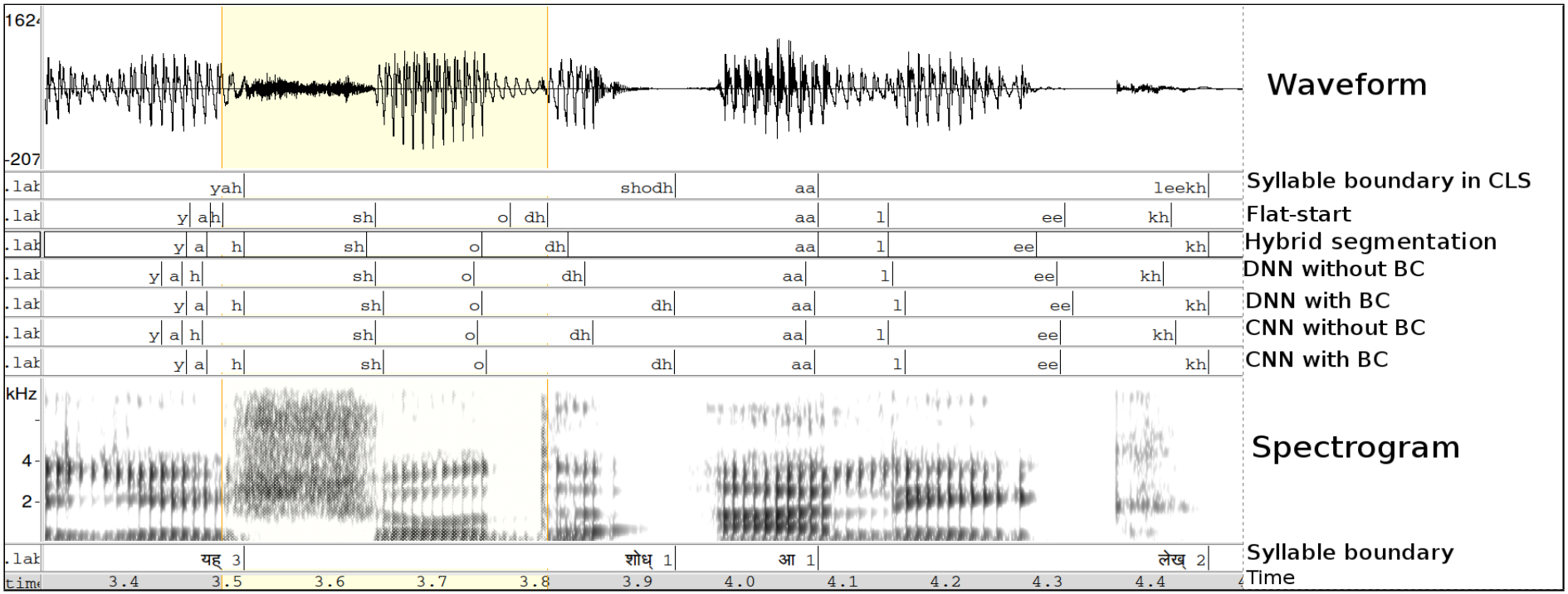


Figure: Example – flat start

Syllable /sh o dh /

Segmentation Correction

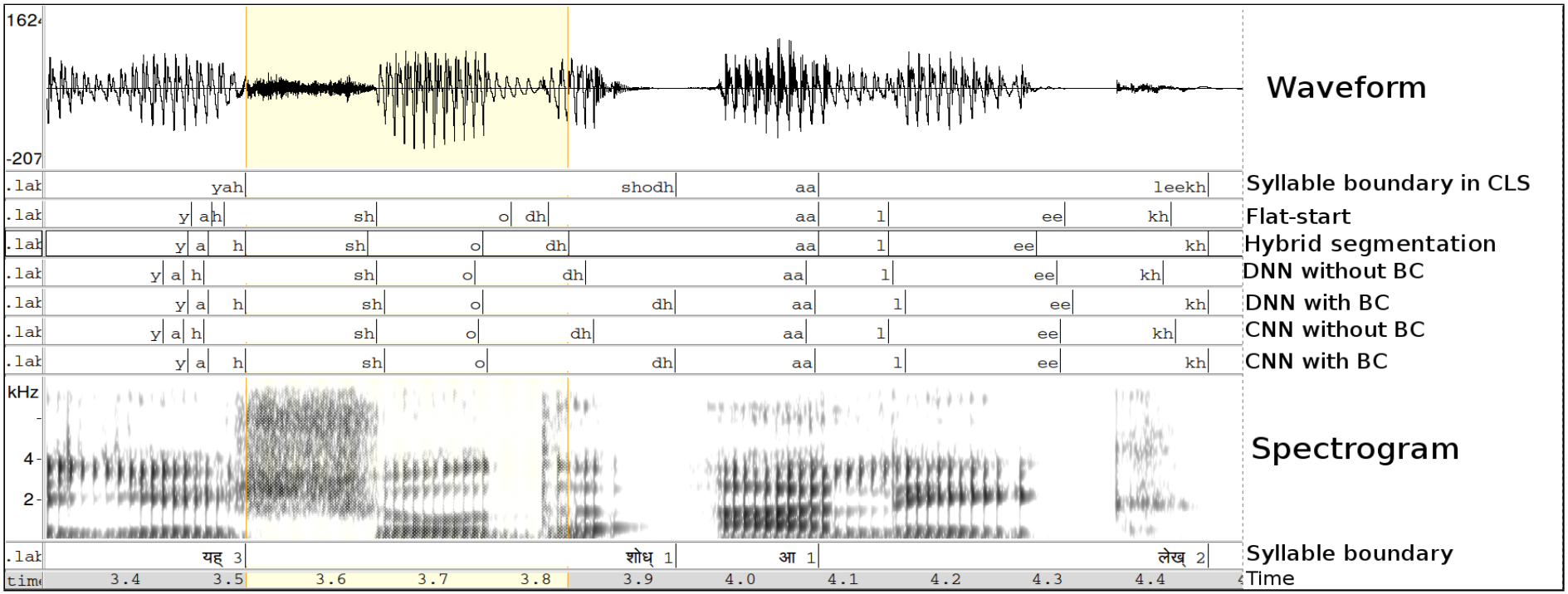


Figure: Example – hybrid segmentation

Syllable /sh o dh /

Segmentation Correction

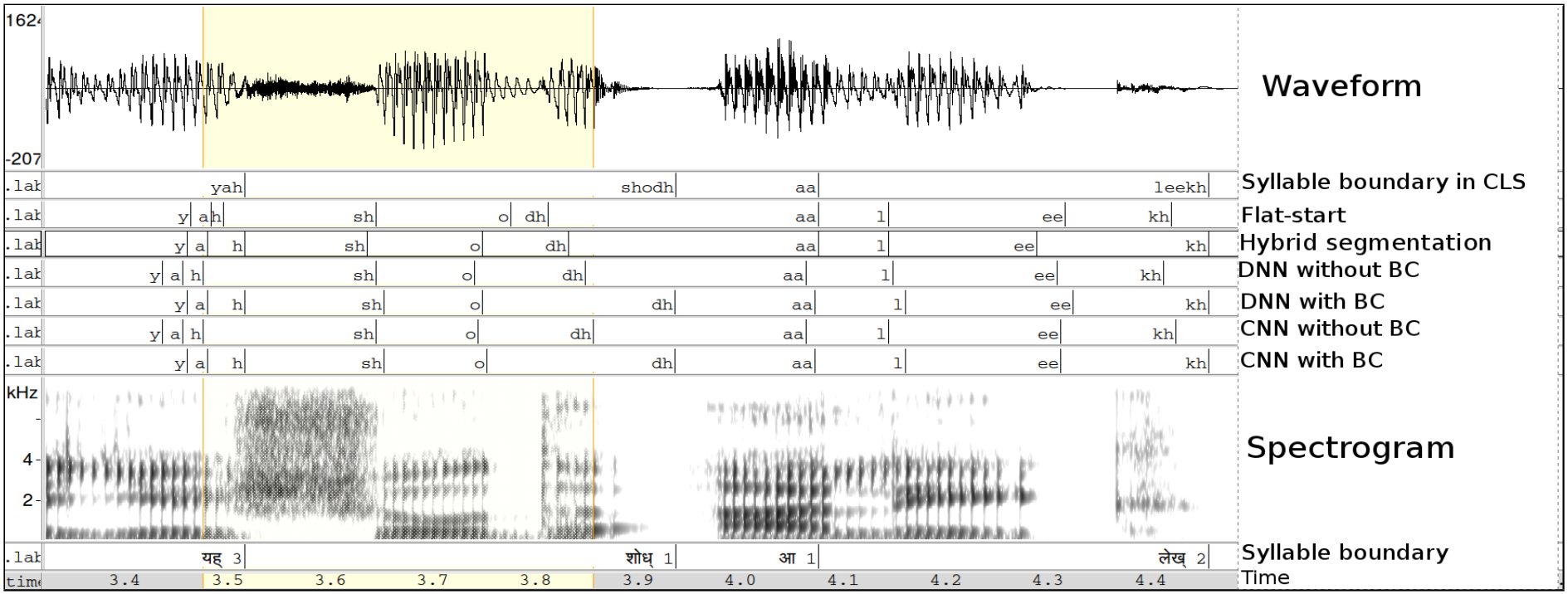


Figure: Example – CNN

Syllable /sh o dh /

Segmentation Correction

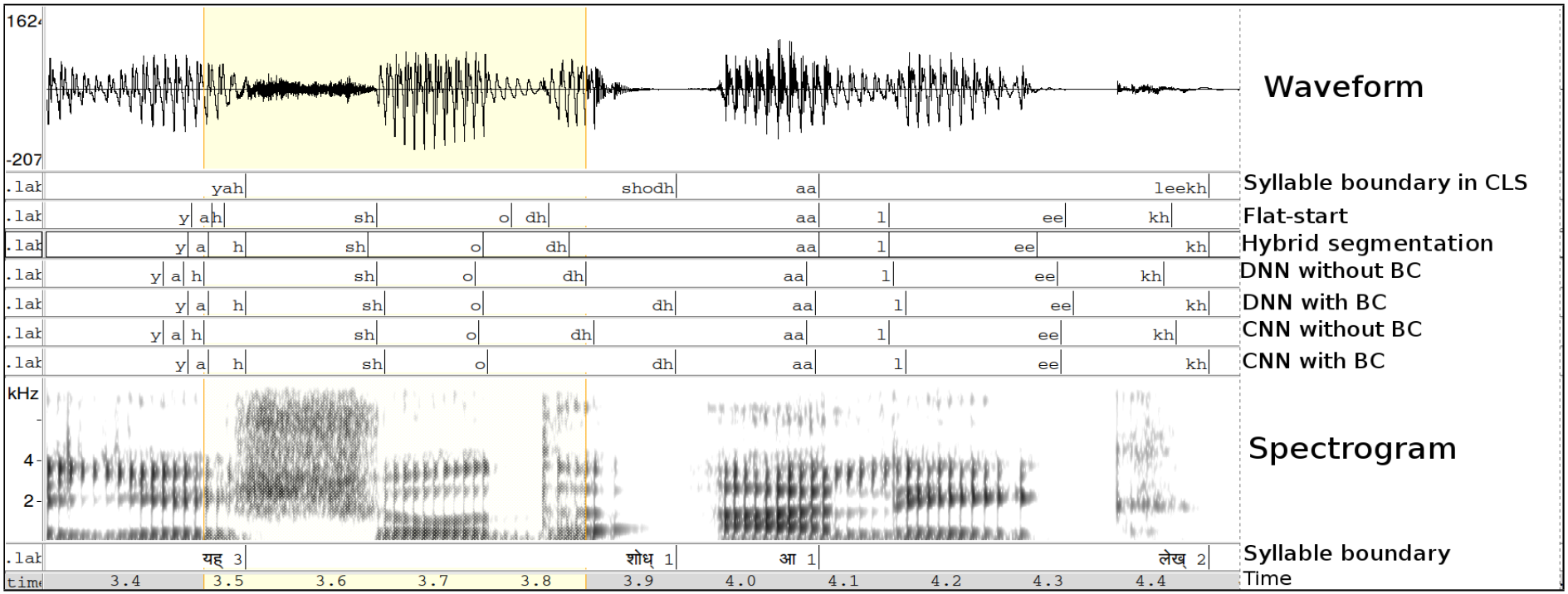


Figure: Example – DNN

Syllable /sh o dh /

Segmentation Correction

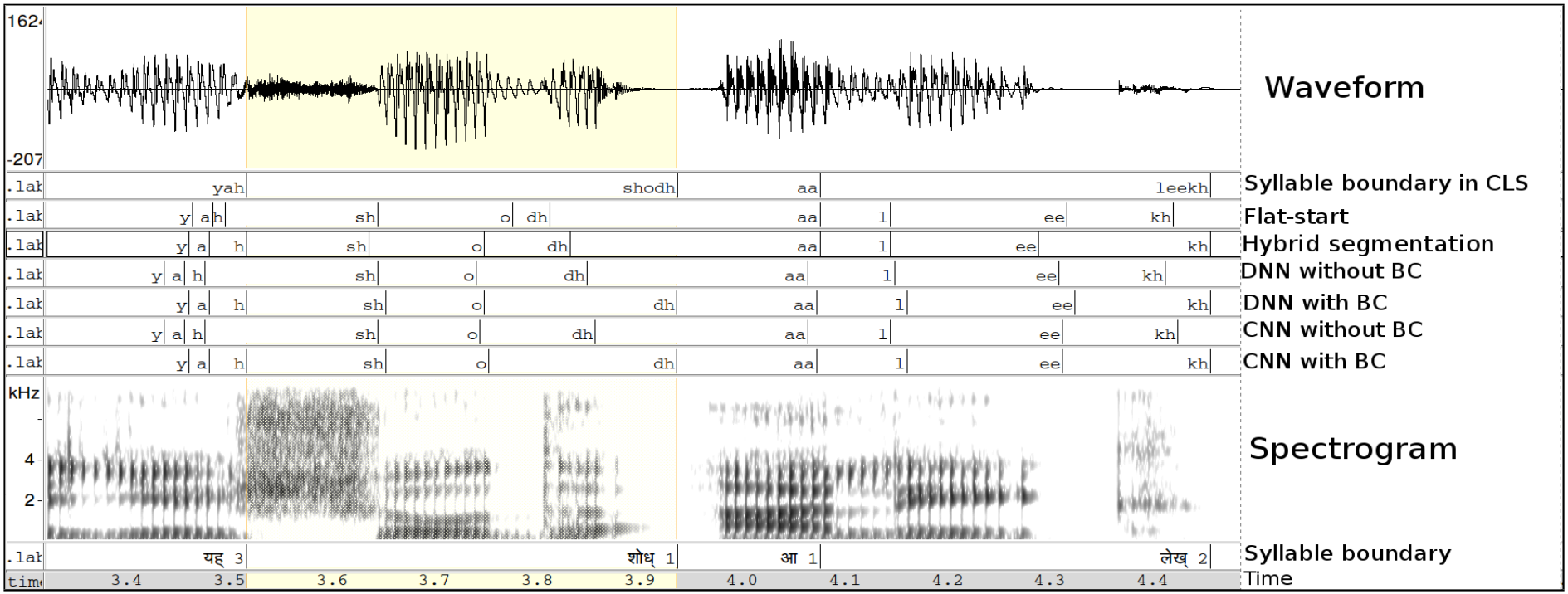


Figure: Example – CNN + BC

Syllable /sh o dh /

Segmentation Correction

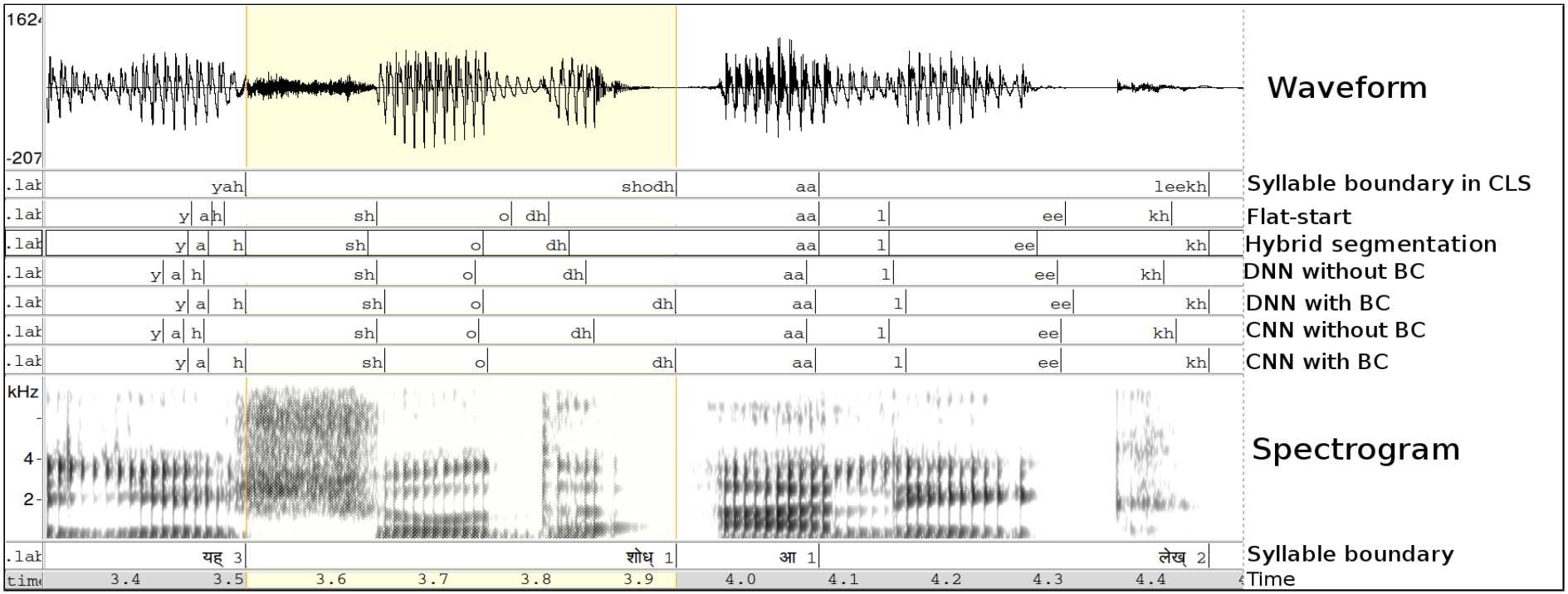


Figure: Example – DNN + BC

Syllable /sh o dh /

Experiments and Results

- A subset of Indic Database is used for the experiments [12]

Table: Dataset used

Language	**Gender**	**Duration** (in hrs)	**No. of utterances**	**No. of phones**
Hindi	Male	5.00	2192	58
Hindi	Female	5.00	2144	58
Bengali	Male	5.00	3093	52
Kannada	Male	3.43	1289	49
Kannada	Female	3.82	1229	48
Malayalam	Male	5.00	3063	52
Telugu	Male	4.24	2478	49

[12] https://www.iitm.ac.in/donlab/tts/database.php

Experiments and Results I

Table: Degradation Mean Opinion Scores (DMOS)

Language	CNN	CNN-BC	DNN	DNN-BC	GMM-BC
Hindi-male	4.03	4.32	4.08	4.55	3.99
Hindi-female	3.35	3.70	3.36	3.51	3.17
Bengali-male	3.26	3.71	3.18	3.60	3.02
Kannada-male	3.64	3.72	3.42	3.44	3.40
Kannada-female	3.13	3.51	3.22	3.44	3.19
Malayalam-male	3.82	4.40	4.02	4.43	3.44
Telugu-male	3.50	4.08	3.67	3.92	3.48

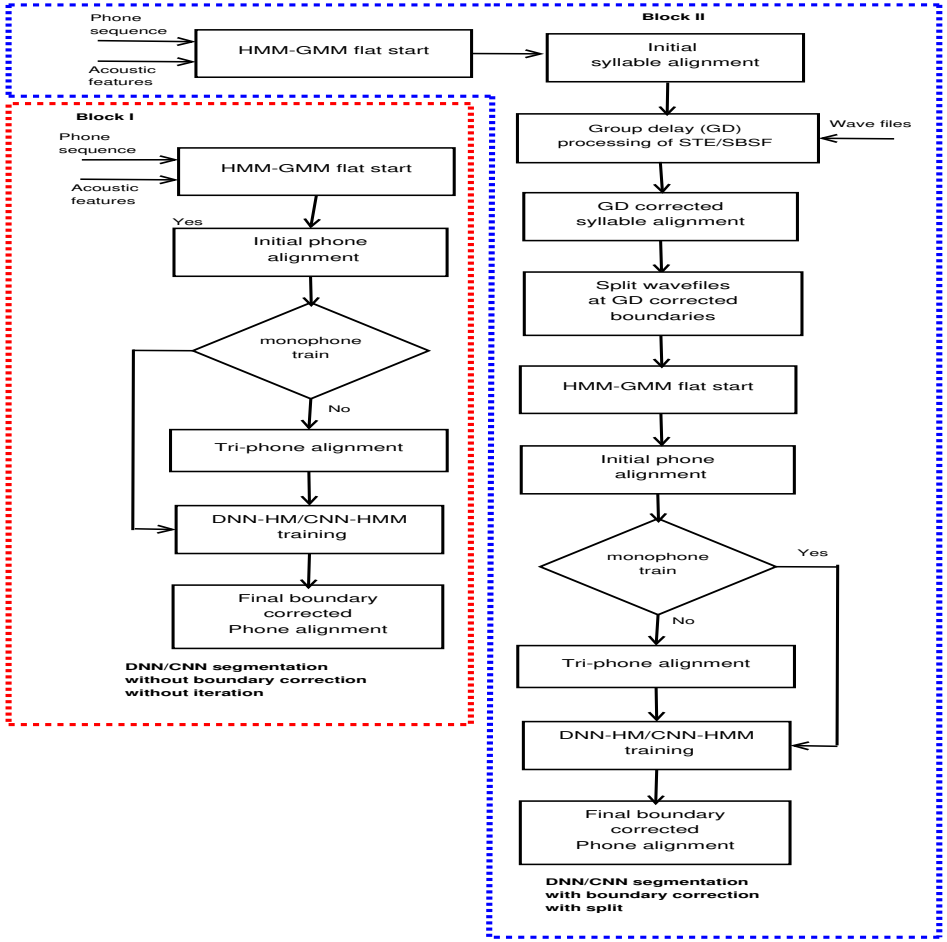
Table: Word Error Rates (%)

Language	CNN	CNN-BC	DNN	DNN-BC	GMM-BC
Hindi-male	3.14	0.28	4.42	2.00	5.85
Hindi-female	7.50	2.50	6.00	1.00	8.75
Bengali-male	6.50	1.81	5.55	1.61	6.40
Kannada-male	4.33	2.00	3.33	2.00	5.66
Kannada-female	4.76	3.57	3.57	2.38	5.95
Malayalam-male	3.33	1.66	3.33	0.50	5.66

Sub-utterance level Approach

Boundary correction at sub-utterance level

Proposed System - Sub-utterance level



Architectures used - Sub-utterance level

- HMM-DNN mono
- HMM-DNN tri
- HMM-DNN mono boundary correction at sub-utterance level
- HMM-DNN tri boundary correction at sub-utterance level
- HMM-CNN mono
- HMM-CNN tri
- HMM-CNN mono boundary correction at sub-utterance level
- HMM-CNN tri boundary correction at sub-utterance level

Experiments and Results

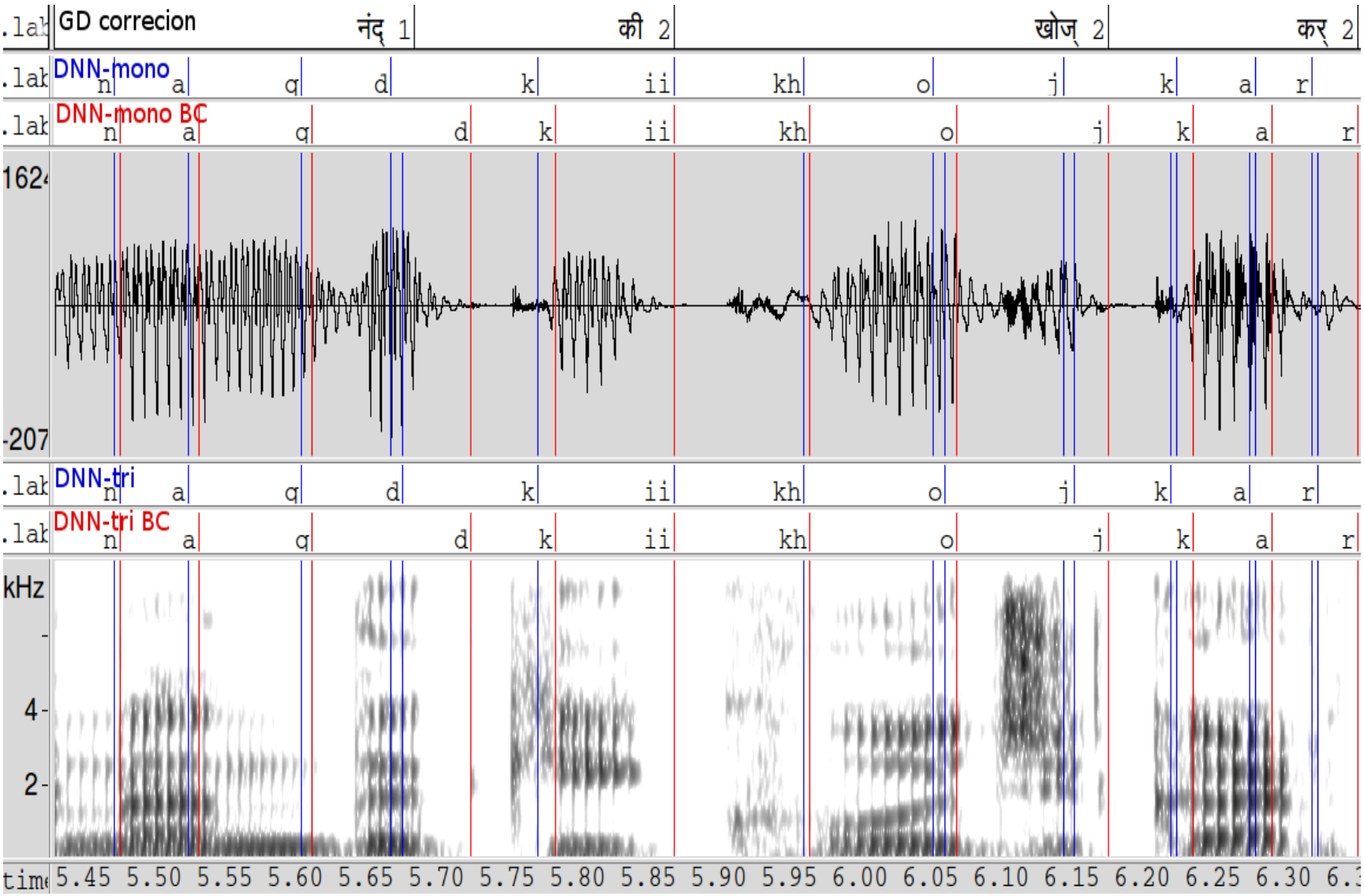


Figure: Sample boundary correction with DNN - Sub-utterance

DNN/CNN Systems

Flat start	Flat start+boundary correction
CNN	CNN+boundary correction
DNN	DNN+boundary correction
Bilingual DNN+BC	Bilingual DNN+BC

Details of this work can be found at IS 2017 [13]

[13] Arun Baby et al, "Deep Learning Techniques in Tandem with Signal Processing Cues for Phonetic Segmentation for Text to Speech Synthesis in Indian Languages," accepted IS 2017

DNN/CNN Systems – Other languages

Bengali	FS+BC	DNN+BC
Bodo	FS+BC	DNN+BC
Gujarati	FS+BC	DNN+BC
Hindi	FS+BC	DNN+BC
Kannada	FS+BC	DNN+BC
Malayalam	FS+BC	DNN+BC
Manipuri	FS+BC	DNN+BC
Marathi	FS+BC	DNN+BC
Rajasthani	FS+BC	DNN+BC
Tamil	FS+BC	DNN+BC
Telugu	FS+BC	DNN+BC

Issues with some of the languages: Erroneous transcription – flagging of the errors using likelihood[14]

[14] Swetha T, Jeena J Prakash and Hema A Murthy,"A semi-automatic method for transcription correction for Indian language TTS systems," NCC 2017.

Scope for Improvement

- The quality is intelligible and more or less natural – naturalness needs improvement.
- Synthesis is still at the sentence level. Synthesis MUST consider prosody across sentences and paragraphs.

Prosody modeling based on sentence structure (NCC 2014) I

Synthesis using syllable position alone

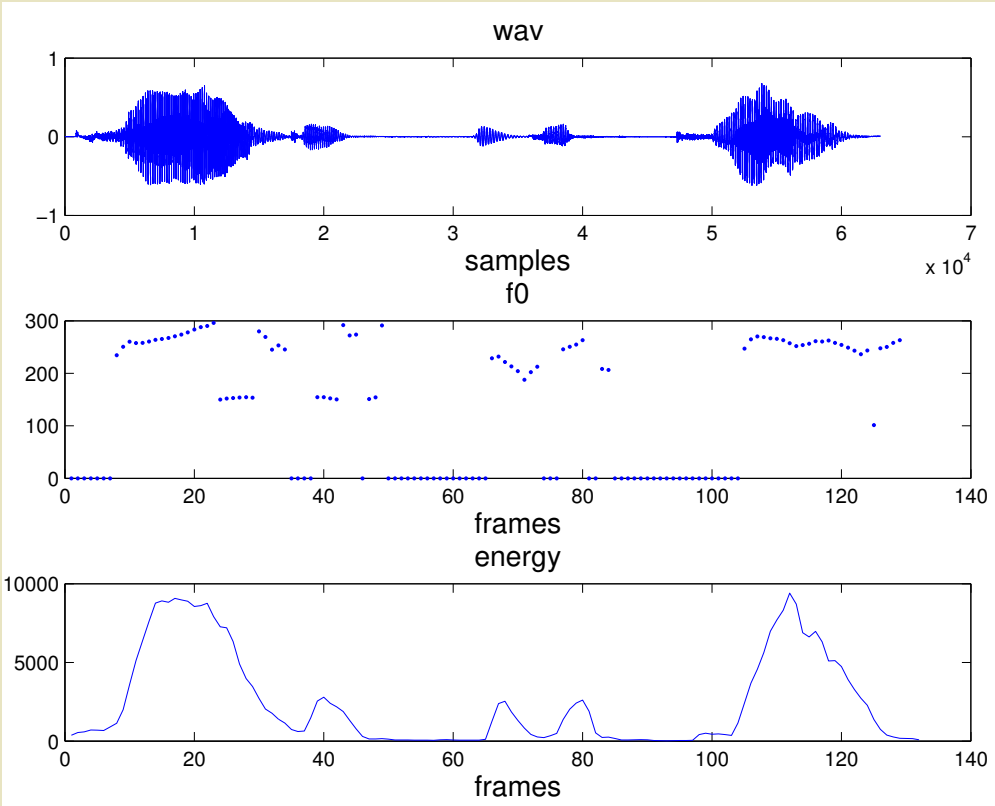


Figure: An example of a speech with artifacts for the text *bhaag bhii khel* which has artifact near *bhii*

Prosody modeling based on sentence structure (NCC 2014) II

Natural sentence prosody

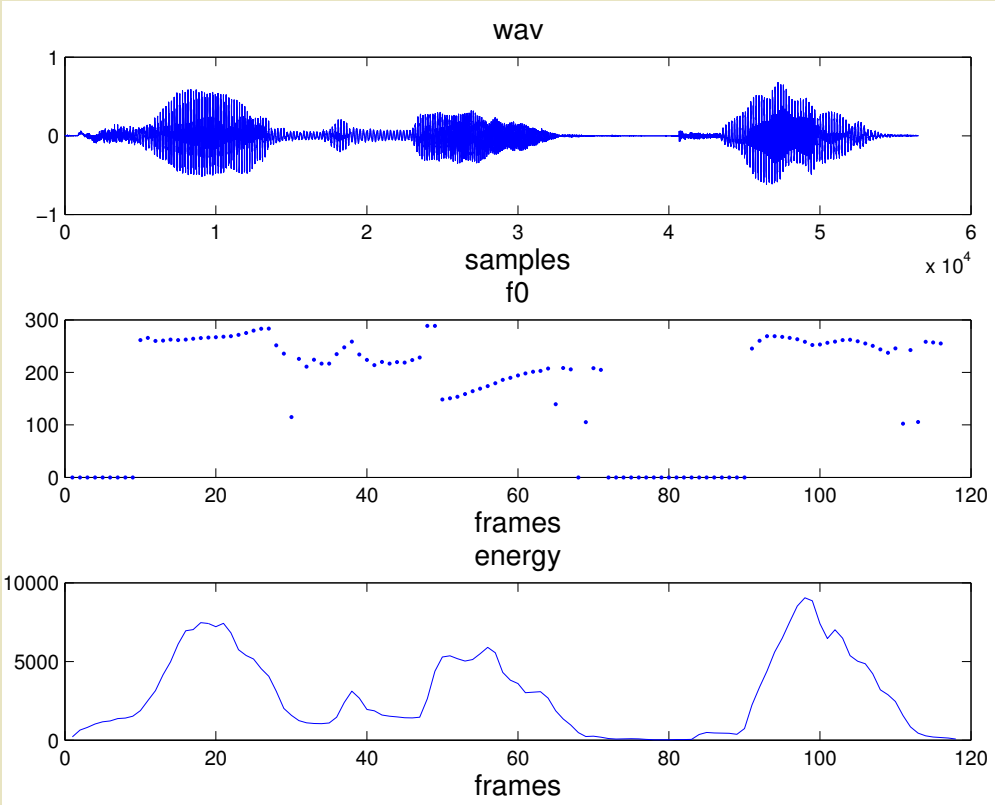


Figure: An example of natural speech for the text *bhaag bhii khel*

Parameters for prosody modeling I

- Intersyllable average f0 difference
- Intersyllable average duration difference
- Intersyllable average energy difference
- Syllables that are part of geminates –e.g /bachcha/ – do not use the syllable /bach/ from this to synthesise /bach/ in /bach/ /pan/.
- likelihood based removal of syllables.

Parameters for prosody modeling II

Modeling duration, f0, energy

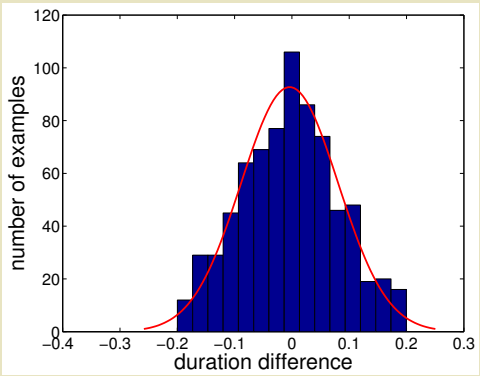


Figure: *

(a)

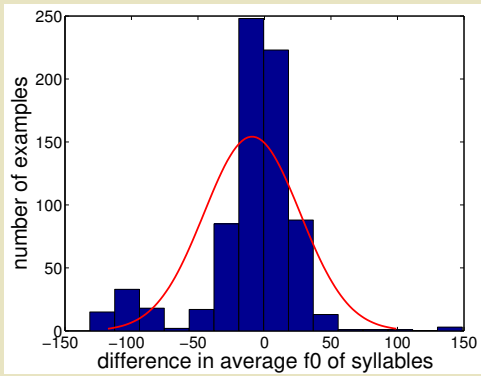


Figure: *

(b)

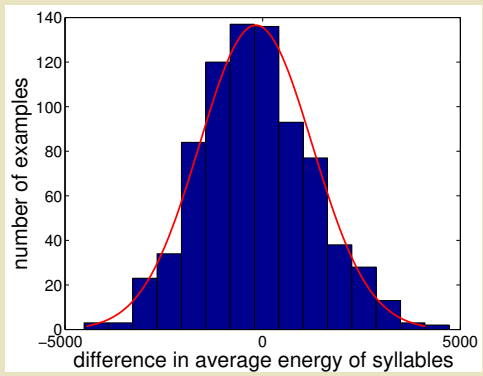


Figure: *

(c)

Figure: (a) Duration difference (degrees of freedom = 14, confidence interval = 0.95), (b) Average pitch difference (degrees of freedom = 17, confidence interval = 0.95), and (c) Average energy difference (degrees of freedom = 25, confidence interval = 0.95)

Example and Results I

Intersyllable prosody based USS

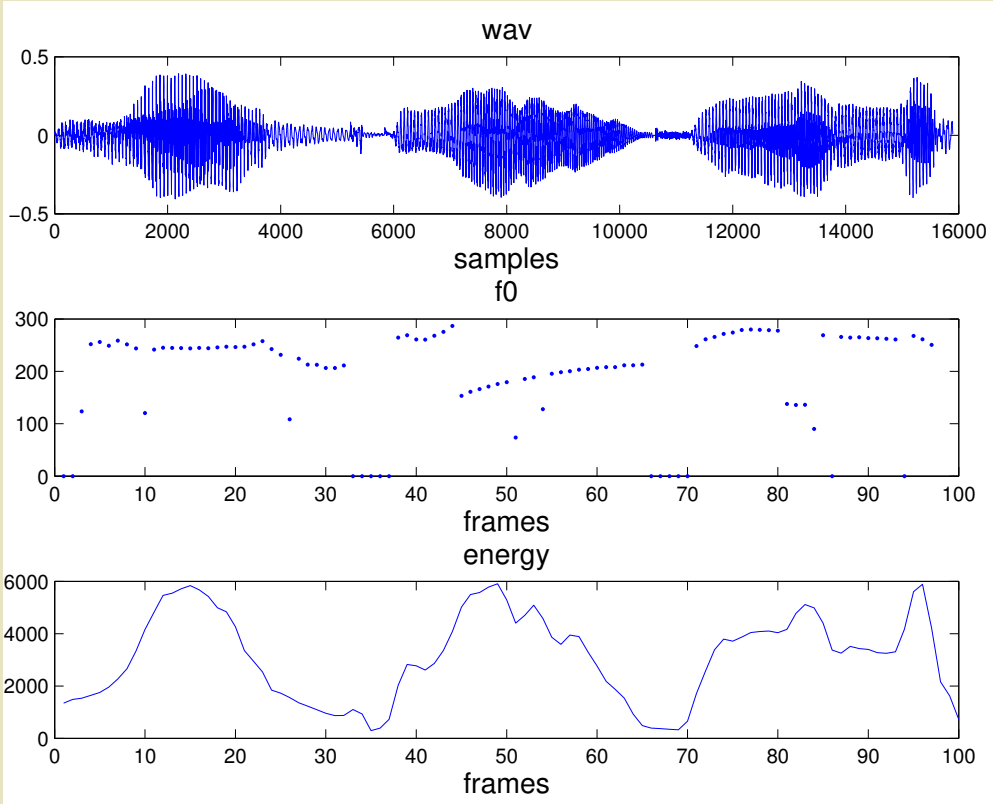


Figure: An example of speech synthesised using the above mentioned approach for the text *bhaag bhii khel*

Example and Results II

DMOS, WER

	DD		f_0 D		ED		DD, ED, f_0 D	
Language	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER
Hindi	3.13	3.03	3.63	5.67	3.24	8.40	3.29	11.94
Tamil	3.10	10.78	3.48	8.97	3.94	1.96	3.27	15.47

Sample Synthesised Sentences

Table: *Synthesised Files*

Hindi Unpruned (USS)	Hindi Pruned (USS)
Tamil Unpruned (USS)	Tamil Pruned (USS)
Hindi HTS	Hindi HTS + STRAIGHT
Hindi HTS+STRAIGHT	Hindi HTS + Prune + STRAIGHT

Other efforts

- Bilingual HTS using common phone set between native language and English.
- Multilingual synthesis (SSNCE, IIITH, IITM)
- Polyglot synthesis (SSNCE, IIITH)
- Replacement of STRAIGHT in HTS (IISc)

Applications

- Highlighting of text on the web
- Online examination for visually challenged
- Small footprint TTS ported to i) Android based systems ii) communication devices like HOPE/KAVI – Chetana (NGO's) products for cerebral palsy patients.
  - Ported to Samsung S5360 Galaxy Y, Samsung Tablet, Adam (notion Ink), Akash Tablet
  - Integration with different Android based products developed at IIT Mandi
  - Integration with OCR (CDAC, Tvm)
  - Integration with ASR for crop information (CDAC, Mumbai)

Awards won and other efforts

- Training Visually Challenged persons on Word Processor, Spreadsheet, e-mail, Internet: 5 years, 180 persons.
- Launch of Screenreader: MoS, MHRD, August 2011.
- Manthan Award: 2012 (Top 74 finalists), 2012.
- GE Innovation and Research Expo Award: 2013.
- Launch of SMS Reader in 5 Indian languages, February 2014.
- Launch of TTS systems in 9 Indian languages on good-goverance day by Hon'ble Minister of IT, Dec 2015.
- Freely available on the web: www.iitm.ac.in/donlab/tts/

Most importantly: Please check our websites

http://www.iitm.ac.in/donlab/hts/ – statistical parametric synthesisers
http://www.iitm.ac.in/donlab/festival/ – unit selection synthesisers

Give your Feedback
Thank you for your Attention

Acknowledgements: Prof. B Yegnanarayana, Prof. K V Madhu Murthy, Prof.C Chandra Sekhar and ALL my students who TAUGHT/TEACH me.

Publicity

Mark your calendars – see you at INTERSPEECH 2018
Sept 2-6, 2018, Hyderabad HICC, India