Text-to-Speech Synthesis

Berkeley
N L P

EECS 183/283a: Natural Language Processing

Text-to-Speech Synthesis

- The artificial conversion of text to speech

alexa

Google

Siri

Hi, I'm Cortana.

- Evaluation :
  - Is it natural?
  - Is it intelligible ?
  - Is it close to the target speaker/style

Physical models for Speech Synthesis

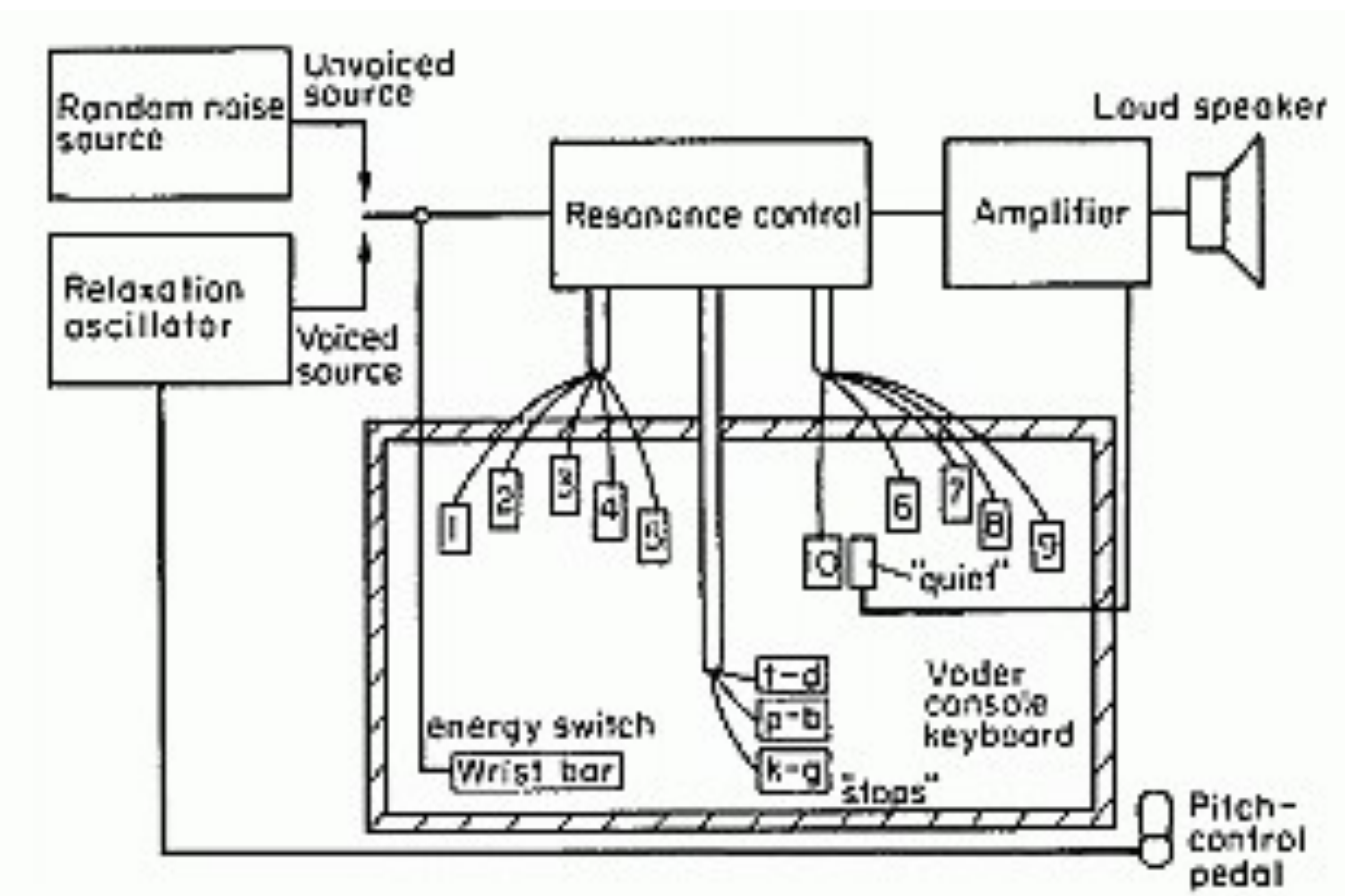
- Blowing air through tubes (1791)
  - Von Kempelen's synthesizer
    - Small whistles for consonants
    - Rubber mouth and nose
    - Bellows provided stream of air

[s]

Electrical models for Speech Synthesis

- Synthesis by electrical simulation
  - Homer Dudley's Voder 1939

[s]



Historical Timeline of TTS

- Formant Synthesis (60s-80s)
  - Waveform construction from components
- Diphone Synthesis (80s-90s)
  - Waveform by concatenation of small number of speech instances
- Unit Selection (90s-00s)
  - Waveform construction from a large number of speech instances
- Statistical Parametric Synthesis (00s-2014)
  - Waveform construction from components
- Neural Synthesis (2014-..)
  - Deep neural networks for speech generation

Formant Synthesis

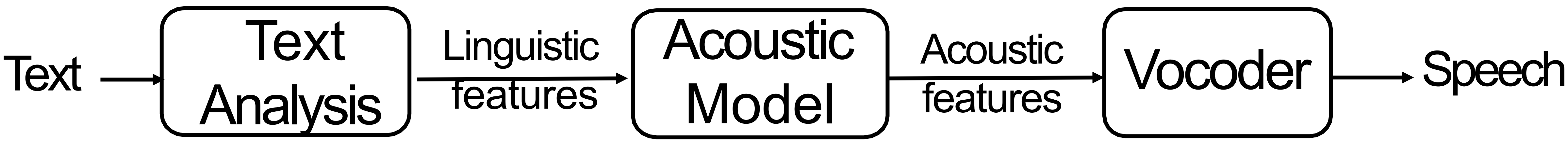
- How does it work?
  - produce speech by mimicking the formant structure and other spectral properties of natural speech
  - using additive synthesis and an acoustic model (with parameters like voicing, fundamental frequency, noise levels)

- Advantages:
  - highly intelligible, even at high speeds
  - well-suited for embedded systems, with limited memory and computation power
- Limitations:
  - not natural, produces artificial, robotic-sounding speech, far from human speech
  - difficult to design rules that specify model parameters

Waveform Synthesis

- Formant synthesis
- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection
- Statistical Parametric Synthesis
- Neural Speech Synthesis

Components of Text-to-Speech Systems



- Popular use case : Text-to-Speech conversion

- Text Analysis
  - Strings of characters to words

- Linguistic Analysis
  - Words to Pronunciation and Prosody

- Waveform Synthesis
  - From pronunciations to waveforms

Text Analysis

- *This is a pen.*
- *My cat who lives dangerously has nine lives.*
- *He stole $100 from the bank.*
- *He stole 1996 cattle on 25 Nov 1996.*
- *He stole $100 million from the bank.*
- *It's 13 St. Andrew St. near the bank.*
- *Its a PIII 1.5Ghz, 512MB RAM, 160Gb SATA, (no IDE) 24x cdrom and 19" LCD.*
- *My home pgae is http://www.geocities.com/awb/.*

What are we trying to solve ?

He retired from business about 1790 with £10,000.

What are we trying to solve ?

He retired from business about
1790 with £10,000.

HE RETIRED FROM BUSINESS ABOUT
SEVENTEEN NINETY WITH TEN THOUSAND POUNDS

What are we trying to solve ?

This should be 14 inches long and
3" by 3" inside, made of hard
wood ¾" thick.

THIS SHOULD BE FOURTEEN INCHES LONG AND
THREE INCHES BY THREE INCHES INSIDE MADE OF HARD
WOOD THREE QUARTERS OF AN INCH THICK

Tokenization

It was almost a matter of course that Dr. Johnson, on arriving in Edinburgh, August 17, 1773, should have come to the White Horse, which was then kept by a person of the name of Boyd.

Non-Standard Words

It was almost a matter of course that Dr. Johnson, on arriving in Edinburgh, August 17, 1773, should have come to the White Horse, which was then kept by a person of the name of Boyd.

NSWs in regular text

- Words not in the lexicon

Text Type	%NSW
Novels	1.5%
Press wire	4.9%
Email	10.7%
Recipes	13.7%
Classifieds	27.9%
IM	20.1%

Ambiguous written forms: homographs

- Homographs
  - Same writing, different pronunciation
  - (Homophones: same pronunciation different writing. “to” “two” “write” “right”)

- English: not many:
- Stress shift (Noun/Verb)
- Segment, project, convict
  - Semantic
    - Bass, read, Begin, bathing, lives, Celtic, wind, Reading, sun, wed, …
    - Roman Numerals

Processing NSWs

- How hard are they?
  - Finding them
  - Identifying them
  - Expanding them
- Current processing techniques
  - Ignored
  - Lexical lookup
  - Hacky hand-written rules
  - (not so) Hacky hand-written rules
  - Statistically train models (and hacky hand written rules)

Homograph Disambiguation

- Same tokens in different contexts
- Identify target homograph
  - E.g. numbers, roman numerals, “St”
- Find instances in large text corpora
- Hand label them with correct answer
- Train a decision tree to predict types

NSW: Roman Numerals

- Roman Numerals as cardinal, ordinals or letters
  - Henry V: Part I Act II Scene XI: Mr X I believe is V I Lenin, and not Charles I.
- Ordinal: Henry V
- Number: Part II
- Letter: Mr X
- Times: 2 X 4 inches
- Word: I am.

NSW models

- *What features help predict class:*
  - *The word form itself*
  - *The word “King” “Queen” “Pope” nearby*
  - *A king/queen/pope name nearby*
  - *Capitalization of nearby words.*
- *class: n(umber) l(etter) r(ex) t(imes)*
- *rex rex_names section_names num_digits p.num_digits, n.num_digits,*

  pp.cap, p.cap, n.cap, nn.cap
-

n	II	0	0	0	11	7	2	3	7	0	0	1	1
n	III	0	0	0	3	4	3	3	5	0	0	1	1
r	VII	1	0	0	4	9	3	3	3	1	1	0	0
n	V	0	0	1	3	1	4	1	2	0	1	0	1

Hard cases

- *Some harder roman numeral cases*
  - *William B. Gates III*
  - *Meet Joe Black II*
  - *The madness of King George III*
  - *He’s a nice chap. I met him last year*

Letters, Abbrevs and Words

- How to pronounces an unknown letter sequence:
- Letters: IBM, CIA, PCMCIA, PhD
- Words: NASA, NATO, RAM
- Abbrev: etc, Pitts, SqH, Pitts Int. Air.
- Hybrids: CDROM, DRAM, WinNT, MacOS
- Letter language model (letter frequencies)

Domain Knowledge

- Modify text processing for the domain:
  - `Smith, Bobbie Q,  3337 St Laurence St, Fort Worth, TX 71611-5484,(817)839-3689`

    `Anderson, W,      445    SycamoreWay NE, Lincoln,NE98125-5108,(212)404-9988`

- Standard Mode
- Address Mode

Sometimes need more than text

- Different context requires different delivery
- What will the weather be like today in Berkeley?
  - It will be **rainy** today in Berkeley.
- When will it be rainy in Berkeley?
  - It will be rainy **today** in Berkeley
- Where will it be rainy today?
  - It will be rainy today in **Berkeley**

Mark-up Languages

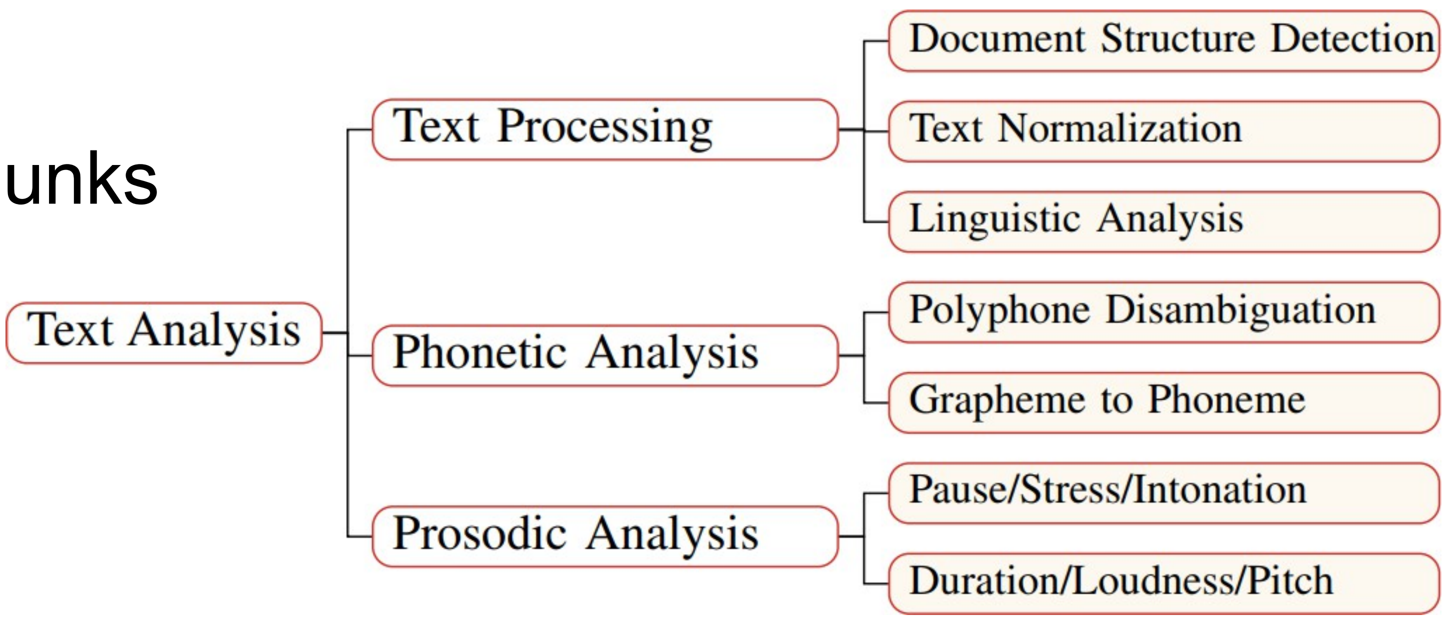
- Add explicit markup to text
- Can be done in machine generated text
- SSML (Speech Synthesis Markup Language)
  - Choice voices, languages
  - Give pronunciations
  - Specifiy breaks, speed, pitch
  - Include external sounds

SSML Example

```
<?xml version="1.0"?>
<!DOCTYPE SABLE PUBLIC "-//SABLE//DTD SABLE speech mark up//EN"
        "Sable.v0_2.dtd"
[]> <SABLE> <SPEAKER NAME="male1">

The boy saw the girl in the park <BREAK/> with the telescope.
The boy saw the girl <BREAK/> in the park with the telescope.

Some English first and then some Spanish.
<LANGUAGE ID="SPANISH">Hola amigos.</LANGUAGE>
<LANGUAGE ID="NEPALI">Namaste</LANGUAGE>

Good morning <BREAK /> My name is Stuart, which is spelled
<RATE SPEED="-40%">
<SAYAS MODE="literal">stuart</SAYAS> </RATE>
though some people pronounce it
<PRON SUB="stoo art">stuart</PRON>.  My telephone number
is <SAYAS MODE="literal">2787</SAYAS>.

I used to work in <PRON SUB="Buckloo">Buccleuch</PRON> Place,
but no one can pronounce that.

By the way, my telephone number is actually
<AUDIO SRC="http://att.com/sounds/touchtone.2.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.7.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.8.au"/>
<AUDIO SRC="http://att.com/sounds/touchtone.7.au"/>.
```

Text Analysis Summary

- Character encodings:
  - Latin-1, iso-8859-1, utf-8 (or special)
- Find tokens
  - White space separated
- Chunk into reasonably sized chunks
  - Sort of sentences
- Map tokens to words
- Disambiguate token types
  - Classify as natural language, or
  - Non-standard Words: abbreviation, numbers, years, money …
- Resolve Ambiguity and find underlying form
- Verbalize NSWs into natural language



Text analysis: linguistic features

- phoneme:
  - current phoneme
  - preceding and succeeding two phonemes
  - position of current phoneme within current syllable
- syllable:
  - numbers of phonemes within preceding, current, and succeeding syllables
  - stress[3] and accent[4] of preceding, current, and succeeding syllables
  - positions of current syllable within current word and phrase
  - numbers of preceding and succeeding stressed syllables within current phrase
  - numbers of preceding and succeeding accented syllables within current phrase
  - number of syllables from previous stressed syllable
  - number of syllables to next stressed syllable
  - number of syllables from previous accented syllable
  - number of syllables to next accented syllable
  - vowel identity within current syllable
- word:
  - guess at part of speech of preceding, current, and succeeding words
  - numbers of syllables within preceding, current, and succeeding words
  - position of current word within current phrase
  - numbers of preceding and succeeding content words within current phrase
  - number of words from previous content word
  - number of words to next content word
- phrase:
  - numbers of syllables within preceding, current, and succeeding phrases
  - position of current phrase in major phrases
  - ToBI endtone of current phrase
- utterance:
  - numbers of syllables, words, and phrases in utterance

Speech Synthesis

- Linguistic Analysis
  - Pronunciations
  - Prosody

Part of Speech Tagging

- Find the most likely tag for each word
  - Finding nouns, verbs, adjectives and others
  - Most words only have one tag (92% correct)
- Context often defines tag type
  - "The project" vs "To project"
- Use HMM Part of Speech tagger
  - But need data to train it (English PennTreeBank)

Poor Man's PoS Tagger

- Hand list "function" word types
  - (determiners a an the this)
  - (conjunctions and or but)
  - (pp in on to)
  - (content everything else)
- Better than nothing
  - Easy to do on new languages

Pronunciation Lexicon

- List of words and their pronunciation
  - (“pencil” n (p eh1 n s ih l))
  - (“table” n (t ey1 b ax l))
- Need the right phoneme set
- Need other information
  - Part of speech
  - Lexical stress
  - Other information (Tone, Lexical accent …)
  - Syllable boundaries

Homograph Representation

- Must distinguish different pronunciations
  - (“project” n (p r aa1 jh eh k t))
  - (“project” v (p r ax jh eh1 k t))
  - (“bass” n_music (b ey1 s))
  - (“bass” n_fish (b ae1 s))
- ASR multiple pronunciations
  - (“route” n (r uw t))
  - (“route(2)” n (r aw t))

Pronunciation of Unknown Words

- How do you pronounce new words
  - 4% of tokens (in news) are new
  - You can't synthesize them without pronunciations
  - You can't recognize them without pronunciations

- Letter-to-Sounds (LTS) rules
  - Grapheme-to-Phoneme (G2P) rules

LTS: Hand written

- Hand written rules
  - [LeftContext] X [RightContext] -> Y
  - e.g.
  - c [h r] -> k
  - c [h] -> ch
  - c [i] -> s
  - c -> k

LTS: Machine Learning Techniques

- Need an existing lexicon
  - Pronunciations: words and phones
  - But different number of letters and phones
- Need an alignment
  - Between letters and phones
  - checked -> ch eh k t

LTS results

- Split lexicon into train/test 90%/10%
  - i.e. every tenth entry is extracted for testing

Lexicon	*Letter Acc*	*Word Acc*
OALD	*95.80%*	*75.56%*
CMUDICT	*91.99%*	*57.80%*
BRULEX	*99.00%*	*93.03%*
DE-CELEX	*98.79%*	*89.38%*
Thai	*95.60%*	*68.76%*

Dialect Lexicons

- Need different lexicons for different dialects
  - US, UK, Indian, Australia, Europeans
- Build dialect independent lexicons
  - Dialect independent vowels (“key-vowels”)
    - The vowel in **coffee** and **conference**
    - Map to aa in US, and o in the UK
  - Post-vocalic r in UK English
    - Car -> k aa
  - Specific words
    - Leisure, route, tortoise, poem

Post-lexical Rules

- Sometimes you need context
- “the” as dh ax or dh iy
- The banana and The apple
- R-insertion in UK English
  - Car door vs car alarm
- Liaison in French
  - Petit vs Petit ami

Linguistic Analysis Summary

- Linguistic analysis
  - Part of speech tagging
  - Pronunciation
    - Phones, stress, (syllables)
    - Letter to sound rules
  - Post lexical rules

Speech Synthesis

- Linguistic Analysis
  - Pronunciations
  - Prosody

Prosody

- How the phonemes will be said
- Four aspects of prosody
  - Phrasing: where the breaks will be
  - Intonation: pitch accents and F0 generation
  - Duration: how long the phonemes will be
  - Power: energy in signal

Phrase Breaks

- Need to take a breath
- Need to chunk relevant parts together
  - Sub-sentential
  - Supra-word
- First approximation
  - At punctuation (comma, semicolon, etc.)
  - Too little
- Second approximation
  - At each (or some) of the content/function words
  - Too much

Phrasing

- Punctuation
  - *Next week, some inmates released early from the Hampton County jail in Springfield, will be wearing a wristband that hooks up to a special jack on their home phones.*

- Content/function words
  - *Next week || some inmates released early || from the Hampton County jail || in Springfield || will be wearing || a wristband || that hooks || up with a special jack || on their home phones.*

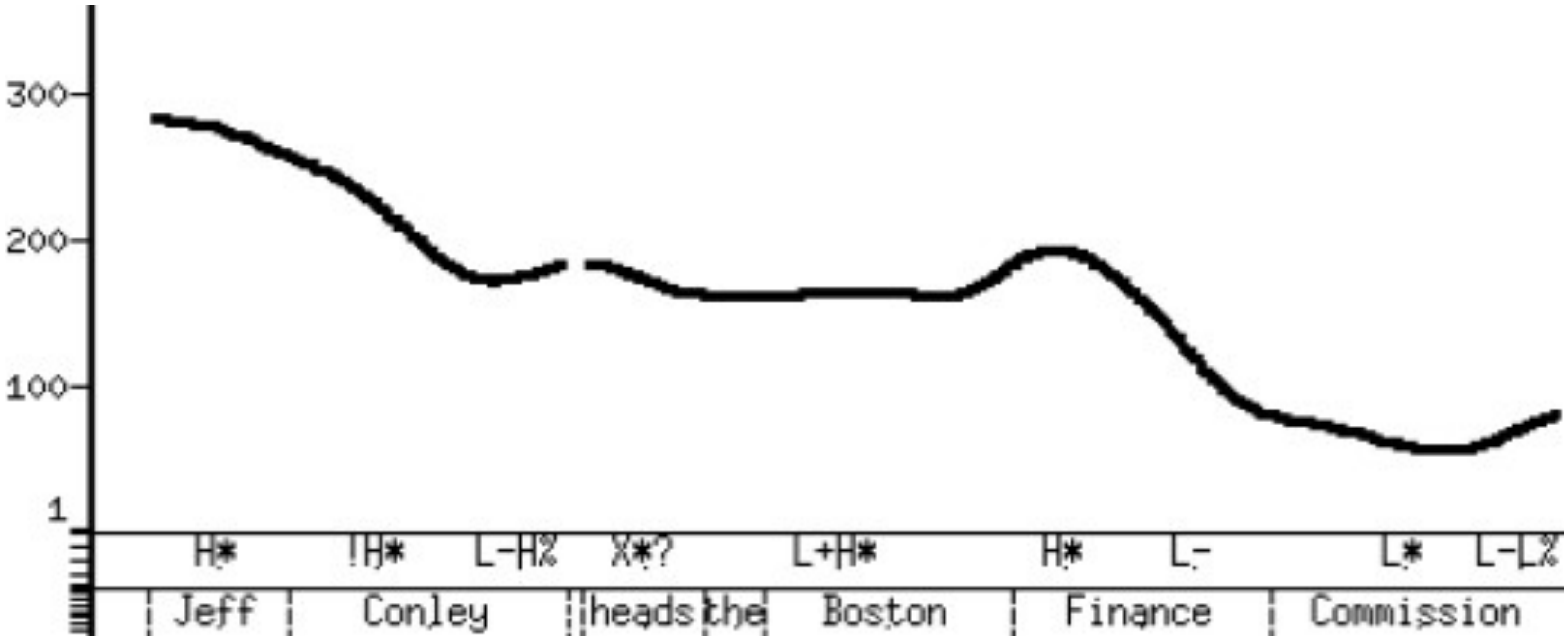
Phrasing

- What is correct?
  - Lots of answers are correct.
  - But some are definitely bad.
- Ostendorf and Vielleux 94
  - Multiple people read same paragraphs
  - If your method matches any single person's version it is correct.

Intonation

- The fundamental tune
  - Accents (highlighting important parts)
  - F0 generation (the tune itself)

Intonation Contour



Intonation Information

- Large pitch range (female)
- Authoritative since goes down at the end
  - News reader
- Emphasis for Finance H*
- Final has a raise – more information to come

- Female American newsreader from WBUR
  - (Boston University Public Radio)

Intonation Examples

- Fixed durations, flat F0.
- Declining F0
- “hat” accents on stressed syllables
- accents and end tones
- statistically trained

Intonational Phonology

- Accents and Boundaries
  - Where are the important changes in F0?
- Accents on syllables
  - Identifies "important" words
    - It will be RAINY today in Boston
    - It will be rainy TODAY in Boston
    - It will BE rainy today IN Boston (strange)

Where do the accents go?

- On important words
- First approximation
  - On stressed syllables in content words
    - It WILL be RAINY TODAY in BOSTON
  - About 80% correct on news reader speech
- ML training can use more features
  - Content, proper nouns, POS, position in text
  - (not semantic information)

ToBI

- Tones and Break Indices
  - A labeling for intonation (English)
- Different accent types
  - H*, !H, L*, L+H*
- Different boundary types
  - L+L%, L+H%, H+H%,

ToBI examples

Marianna made the marmelade.

H*		H*	L-L	default reading
H*			L-L%	emphasis on Marianna
L+H*			L-L%	contrastive reading
L*			H-H%	incredulous
L*		L*	H-H%	doubly incredulous
L+H*L-H%	L*	H*	L-L%	(2 intonation phrases)

F0 Generation

- Contour from accents (and durations)
- Piece together shapes of different accents
- Generated
  - By rule
  - Trained from data

Duration Prediction

- Each phone needs a duration
  - Make it 80ms
- Vowels are typically longer than consonants
- Emphasis/accent/stress lengthens them
- Initial and final phones are longer

Duration PredictionModels

- By rule
  - Klatt rules
- By ML training (using features)
  - linear regression
  - Easy to get reasonable durations
  - Hard to get very good durations

Fast and Slow Speech

- Speaking fast: not uniformly shorter durations
  - Have less prosodic breaks
  - Reduce syllables
  - Make consonants shorter
  - Make vowels a little shorter

1.0x 1.2x

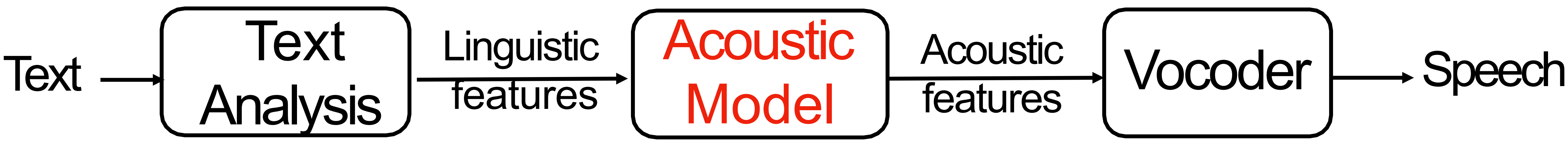
- Speaking slow: not uniformly longer durations
  - Add more prosodic breaks
  - Small increases in vowel duration (?)

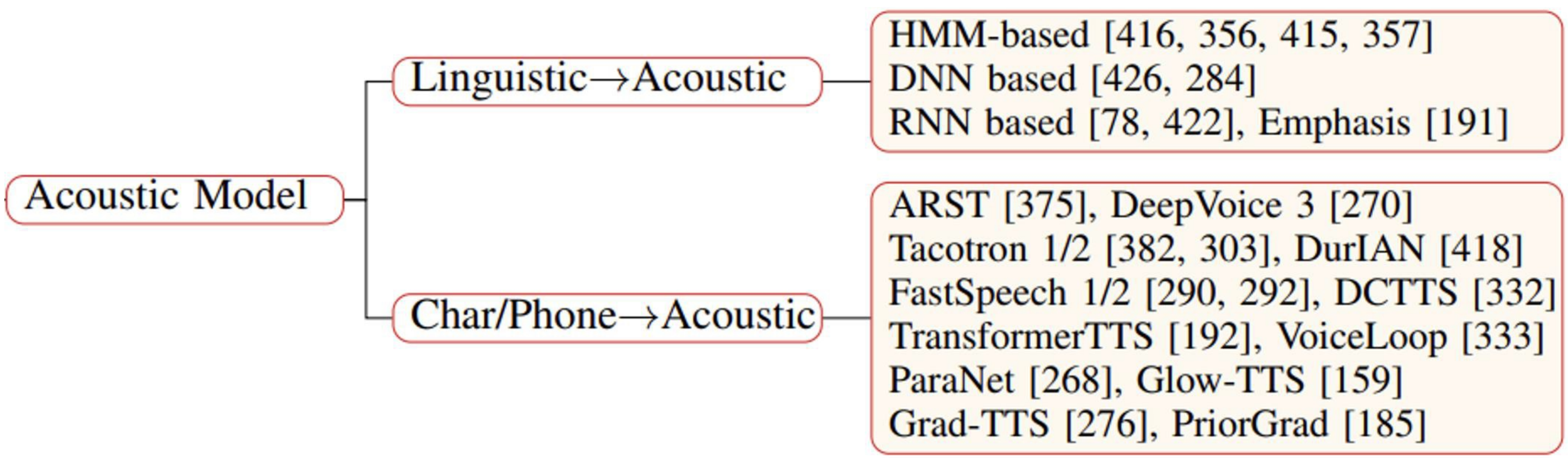
Prosody Summary

- Prosody
  - Phrasing
  - Intonation
    - Accents + F0 generation
  - Duration
  - Power

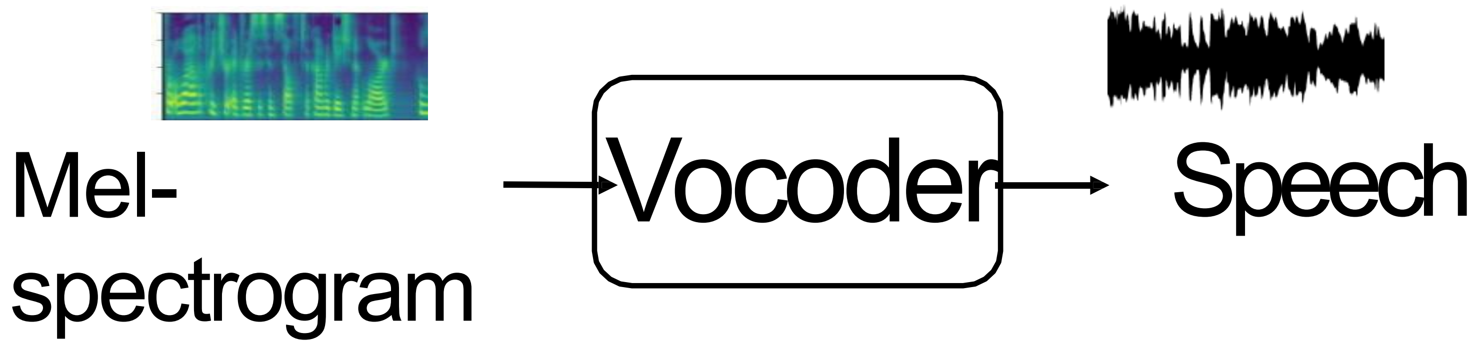
Acoustic model

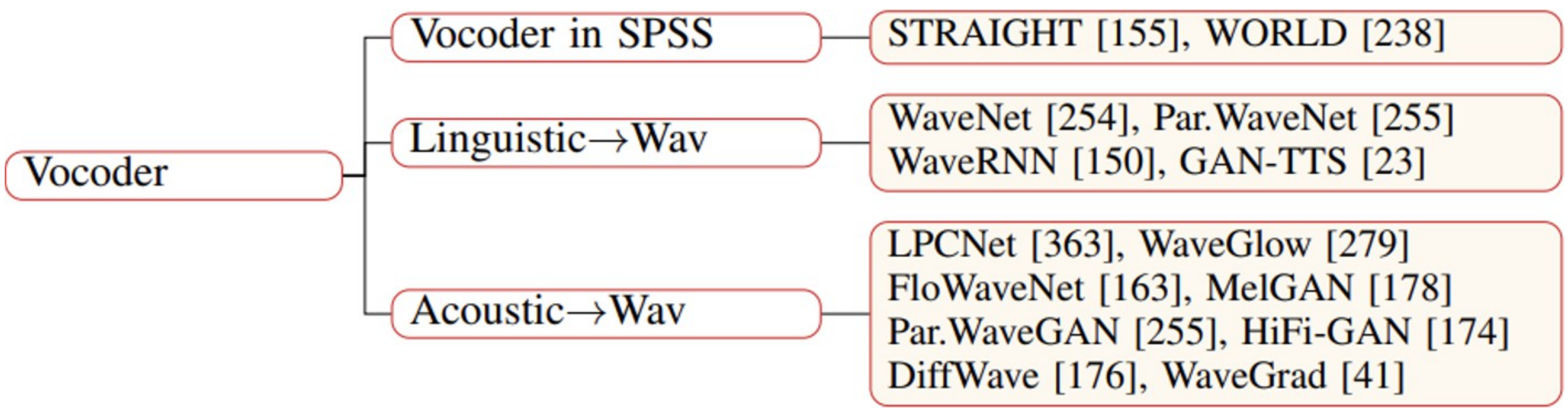
- Predict acoustic features from linguistic features



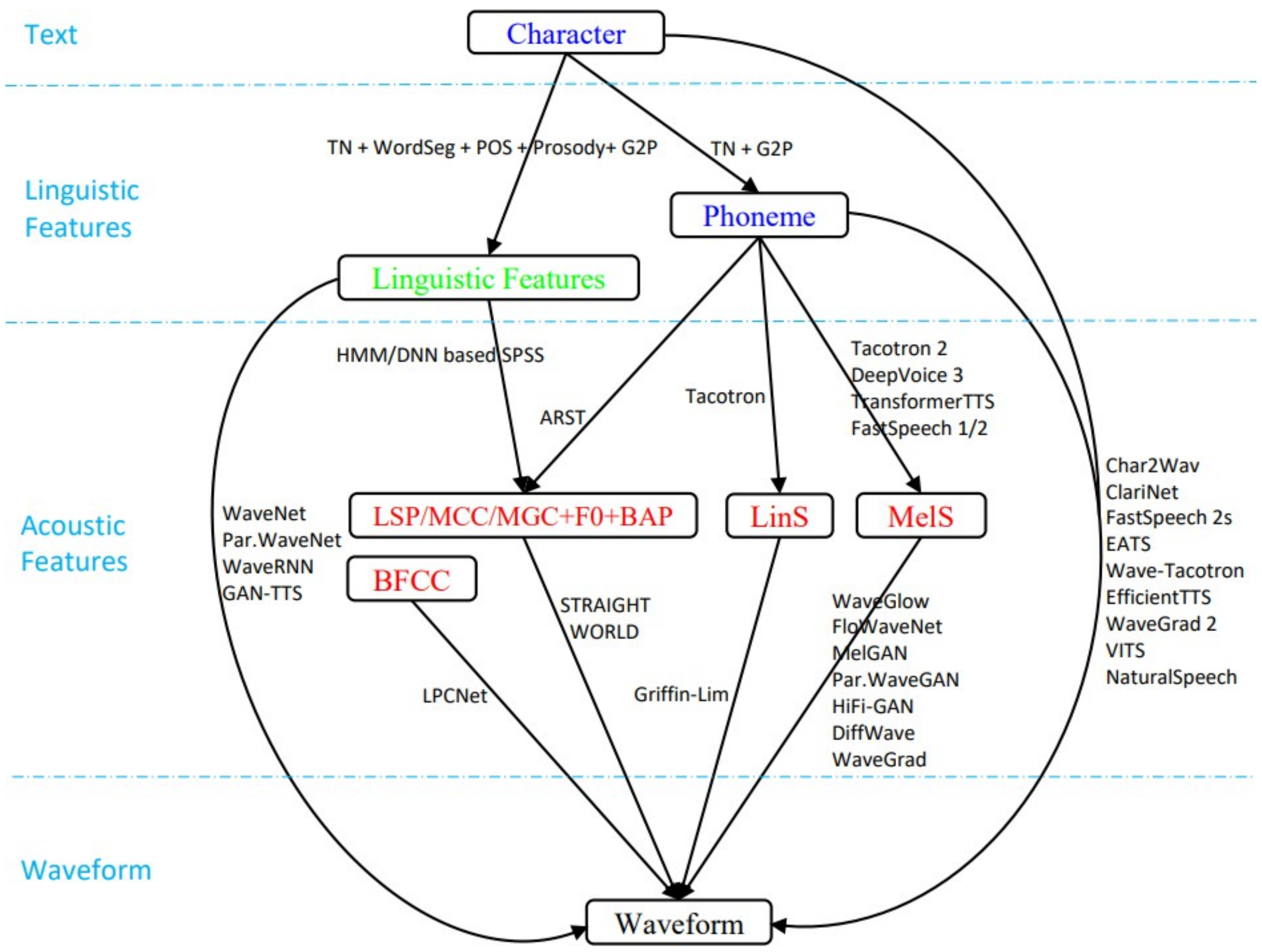


Vocoder

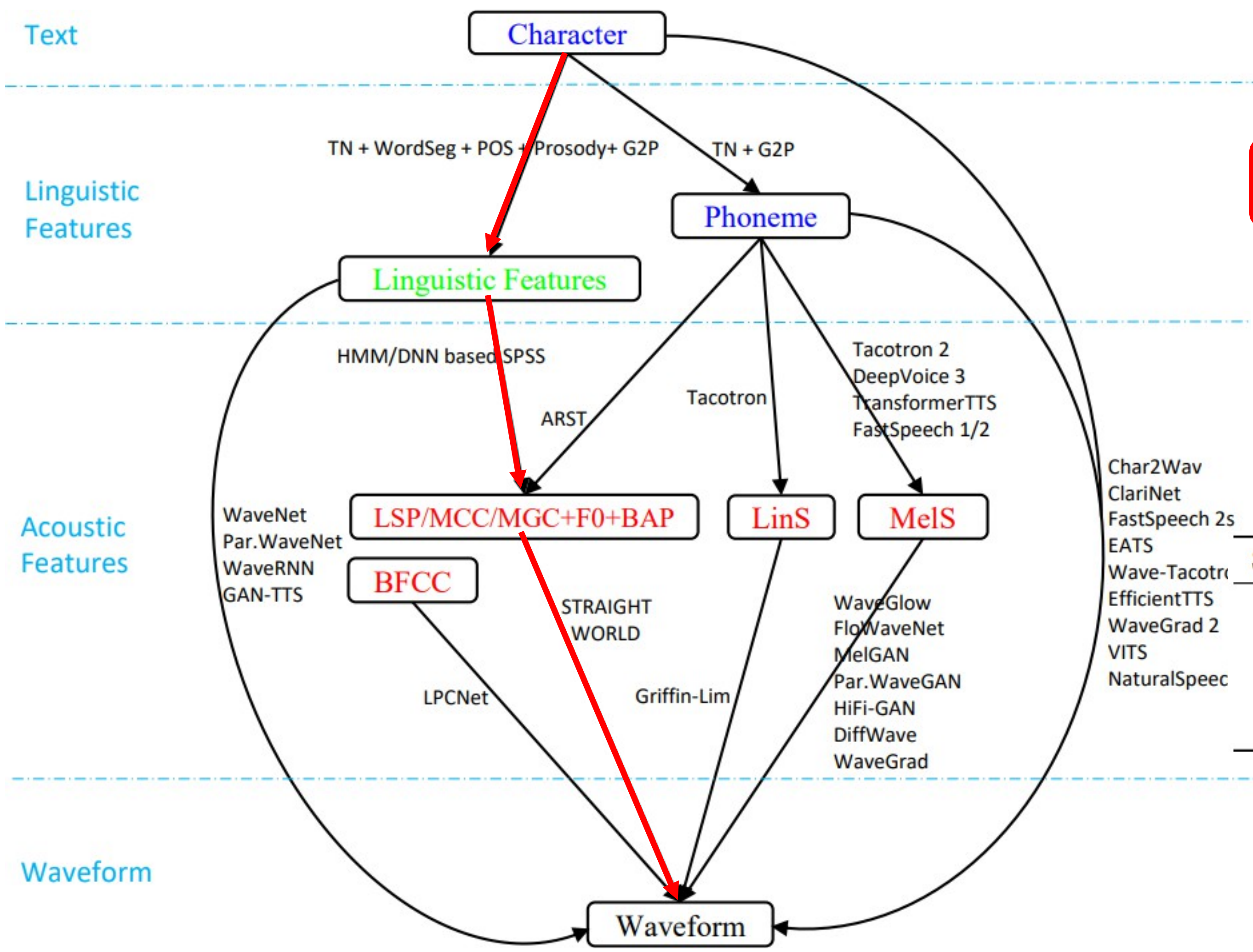

Mel-spectrogram
Vocoder
Speech

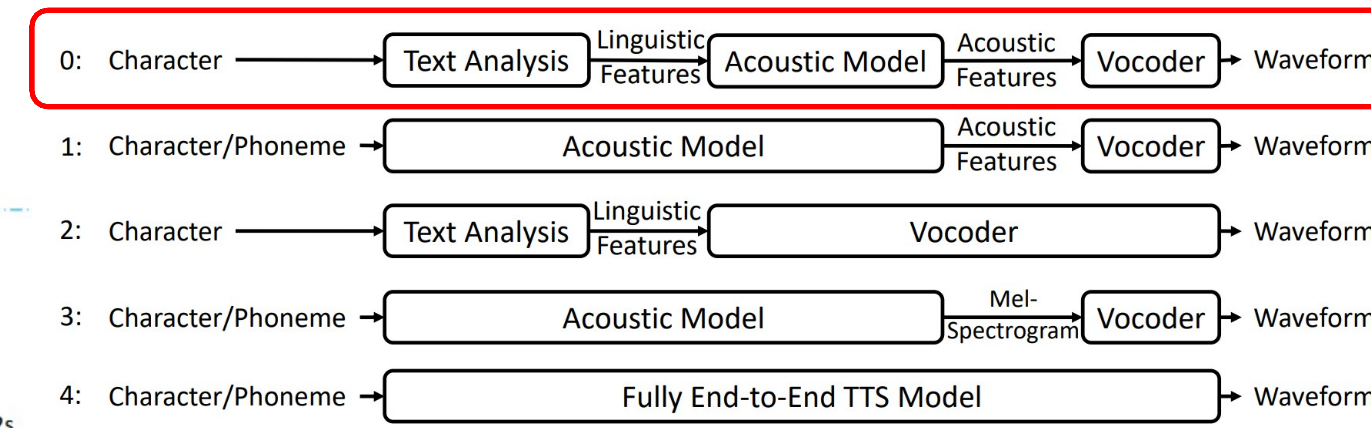

Vocoder
Vocoder in SPSS
STRAIGHT [155], WORLD [238]
Linguistic→Wav
WaveNet [254], Par.WaveNet [255]
WaveRNN [150], GAN-TTS [23]
Acoustic→Wav
LPCNet [363], WaveGlow [279]
FloWaveNet [163], MelGAN [178]
Par.WaveGAN [255], HiFi-GAN [174]
DiffWave [176], WaveGrad [41]

Data conversion pipeline



Data conversion pipeline

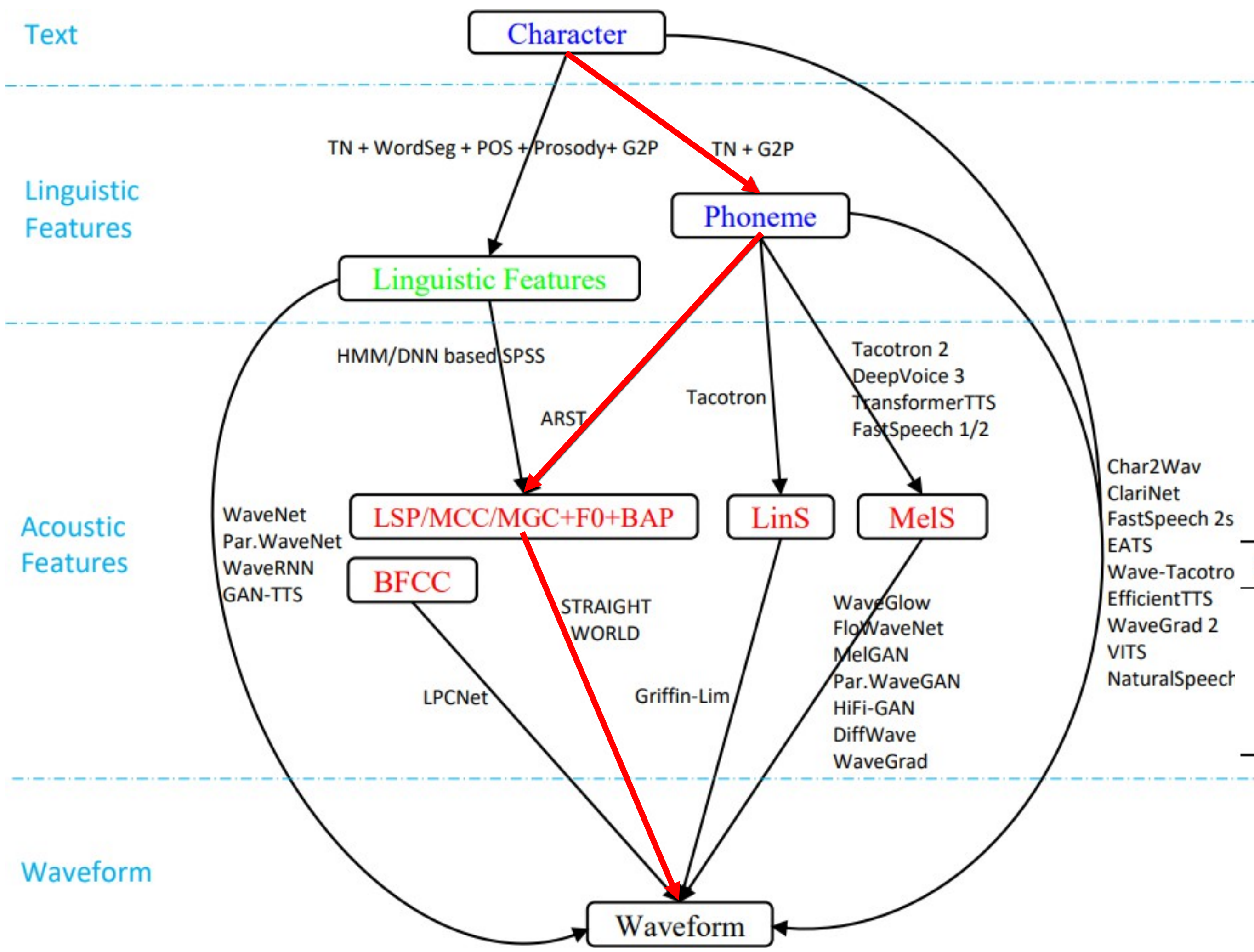


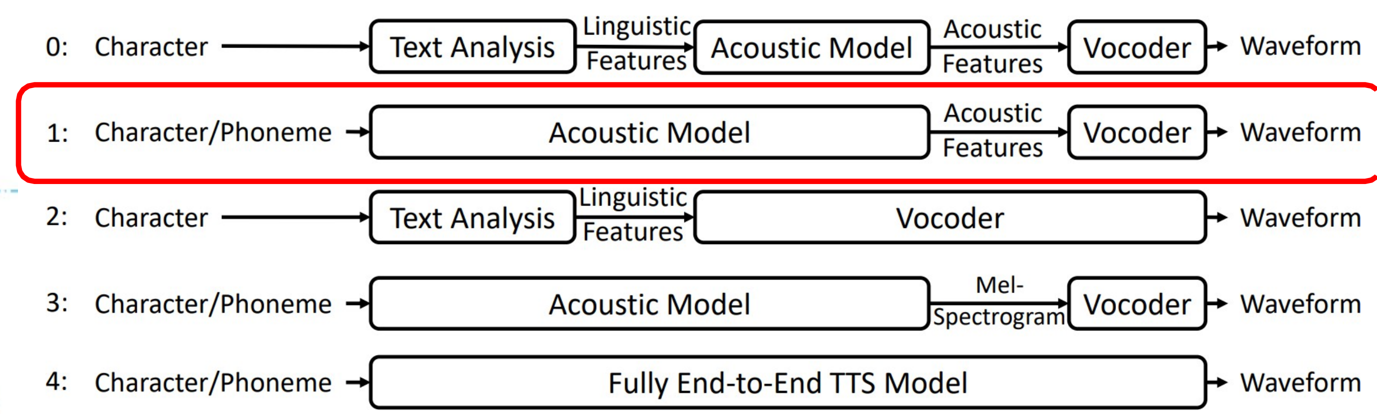


Stage	Models
0	SPSS [418, 358, 417, 427, 359]
1	ARST [377]
2	WaveNet [255], DeepVoice 1/2 [8, 88], Par. WaveNet [256], WaveRNN [151], HiFi-GAN [23]
3	DeepVoice 3 [271], Tacotron 2 [304], FastSpeech 1/2 [291, 293], WaveGlow [280], FloWaveNet [164]
4	Char2Wav [316], ClariNet [270], FastSpeech 2s [293], EATS [70], VITS [161], NaturalSpeech [346]

Path 0

Data conversion pipeline

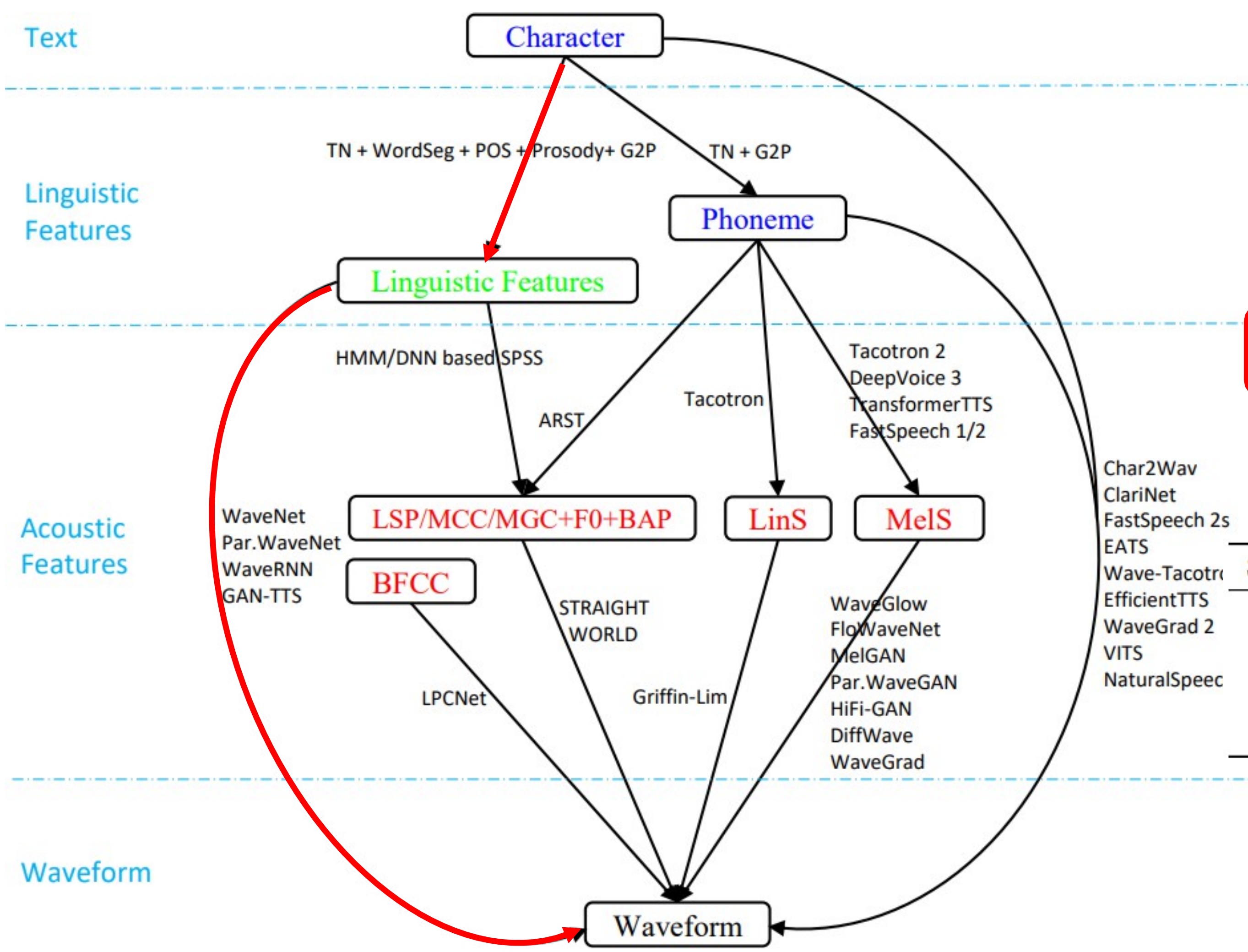


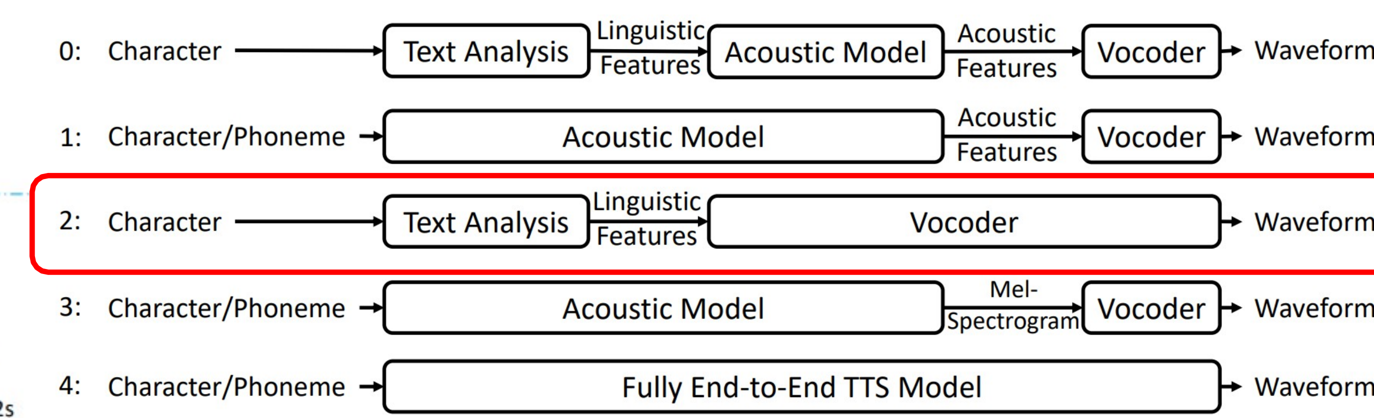


Stage	Models
0	SPSS [418, 358, 417, 427, 359]
1	ARST [377]
2	WaveNet [255], DeepVoice 1/2 [8, 88], Par. WaveNet [256], WaveRNN [151], HiFi-GAN [23]
3	DeepVoice 3 [271], Tacotron 2 [304], FastSpeech 1/2 [291, 293], WaveGlow [280], FloWaveNet [164]
4	Char2Wav [316], ClariNet [270], FastSpeech 2s [293], EATS [70], VITS [161], NaturalSpeech [346]

Path 0

Data conversion pipeline

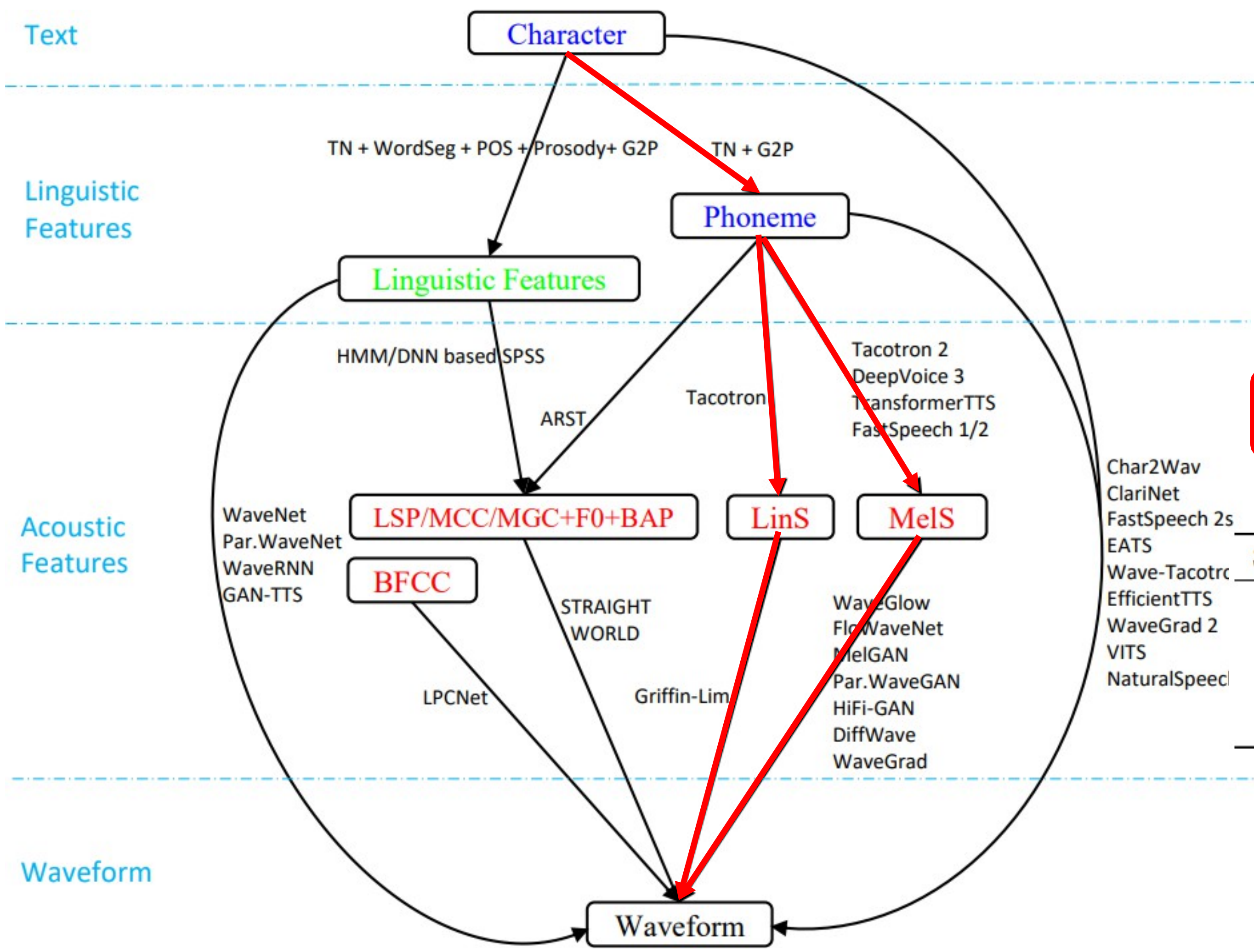


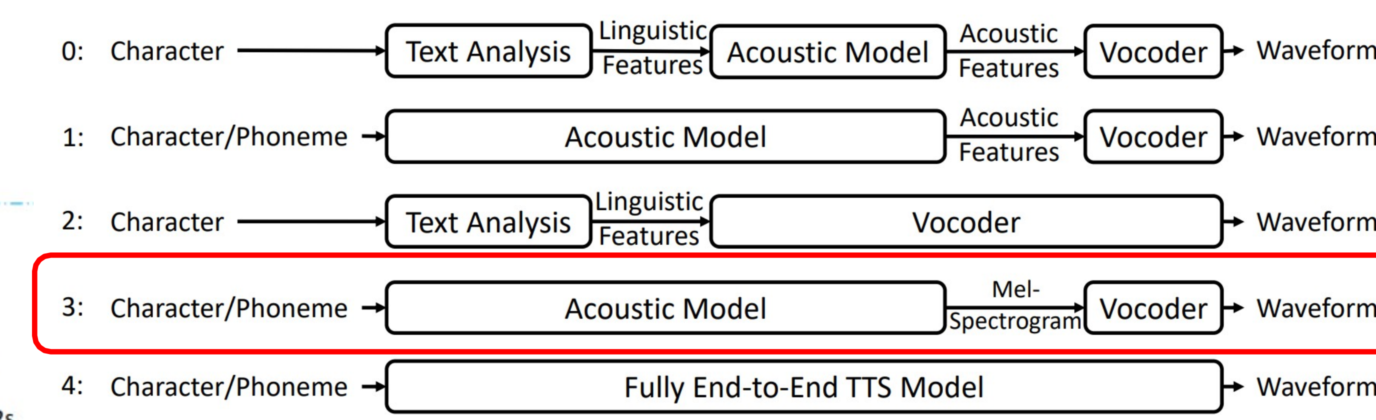


Stage	Models
0	SPSS [418, 358, 417, 427, 359]
1	ARST [377]
2	WaveNet [255], DeepVoice 1/2 [8, 88], Par. WaveNet [256], WaveRNN [151], HiFi-GAN [23]
3	DeepVoice 3 [271], Tacotron 2 [304], FastSpeech 1/2 [291, 293], WaveGlow [280], FloWaveNet [164]
4	Char2Wav [316], ClariNet [270], FastSpeech 2s [293], EATS [70], VITS [161], NaturalSpeech [346]

Path 2

Data conversion pipeline

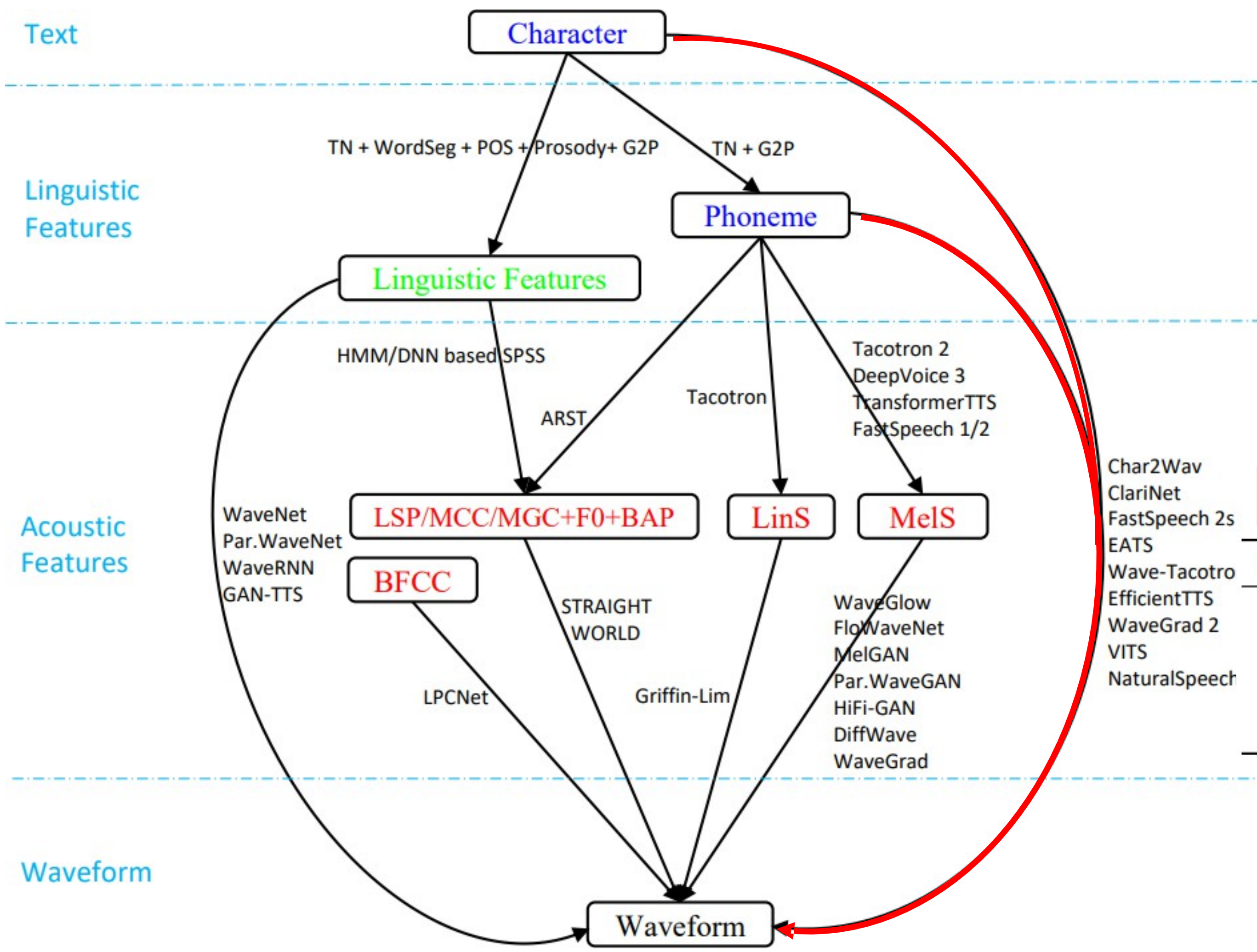




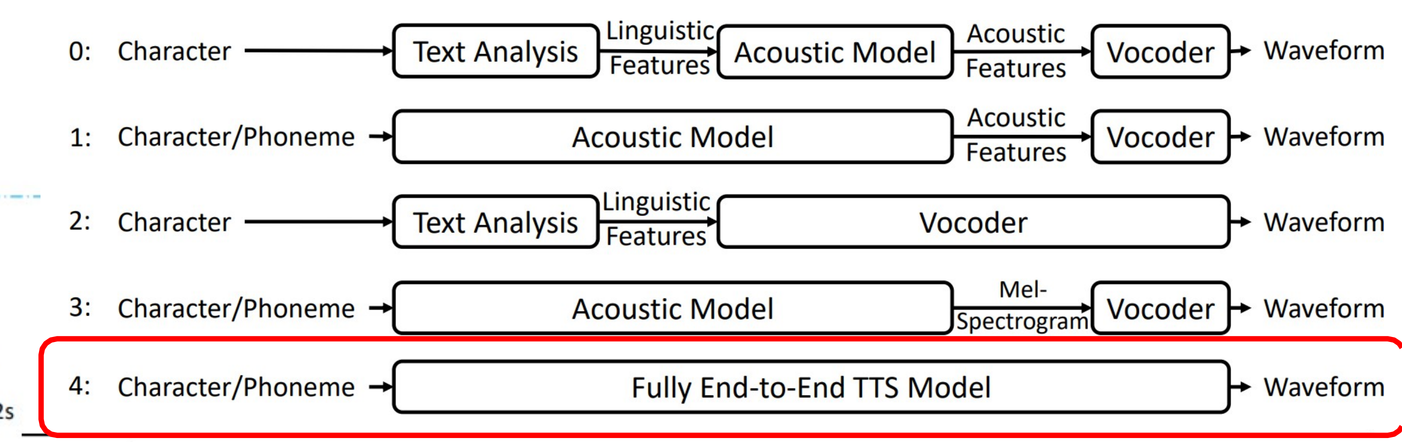
Stage	Models
0	SPSS [418, 358, 417, 427, 359]
1	ARST [377]
2	WaveNet [255], DeepVoice 1/2 [8, 88], Par. WaveNet [256], WaveRNN [151], HiFi-GAN [23]
3	DeepVoice 3 [271], Tacotron 2 [304], FastSpeech 1/2 [291, 293], WaveGlow [280], FloWaveNet [164]
4	Char2Wav [316], ClariNet [270], FastSpeech 2s [293], EATS [70], VITS [161], NaturalSpeech [346]

Path 3

Data conversion pipeline

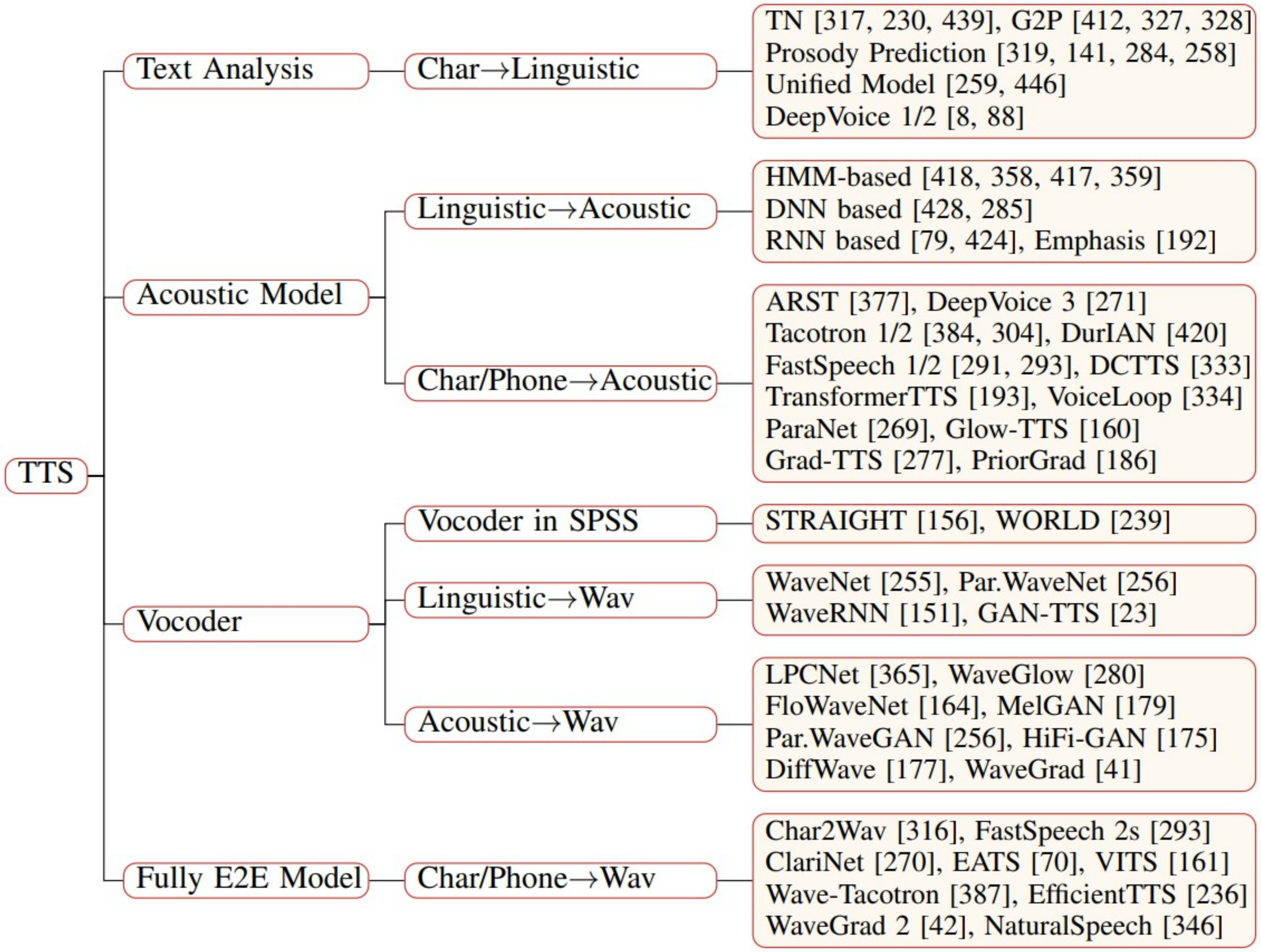


Path 4



Stage	Models
0	SPSS [418, 358, 417, 427, 359]
1	ARST [377]
2	WaveNet [255], DeepVoice 1/2 [8, 88], Par. WaveNet [256], WaveRNN [151], HiFi-GAN [23]
3	DeepVoice 3 [271], Tacotron 2 [304], FastSpeech 1/2 [291, 293], WaveGlow [280], FloWaveNet [164]
4	Char2Wav [316], ClariNet [270], FastSpeech 2s [293], EATS [70], VITS [161], NaturalSpeech [346]

Key components in TTS



2022

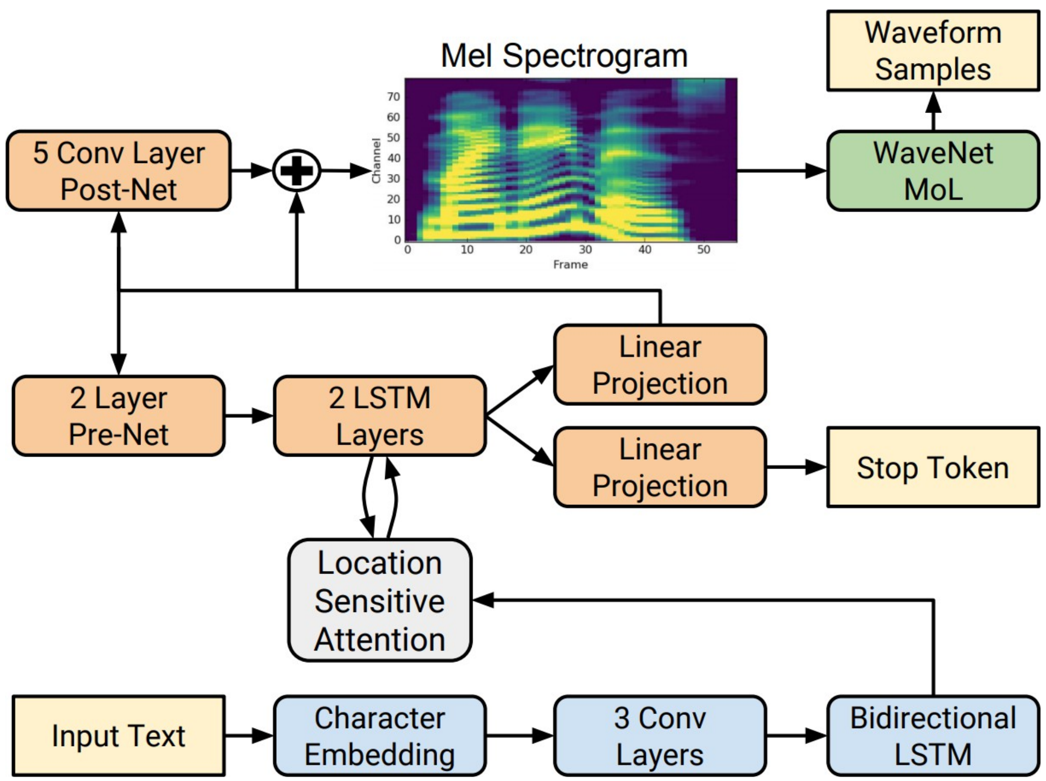
Acoustic model

- Acoustic model in SPSS
- Acoustic models in end-to-end TTS
  - RNN-based (e.g., Tacotron series)
  - CNN-based (e.g., DeepVoice series)
  - Transformer-based (e.g., FastSpeech series)
  - Other (e.g., Flow, GAN, VAE, Diffusion)

Acoustic Model	Input→Output	AR/NAR	Modeling	Structure
SPSS				
HMM-based [424, 363]	Ling→MCC+F0	/	/	HMM
DNN-based [434]	Ling→MCC+BAP+F0	NAR	/	DNN
LSTM-based [79]	Ling→LSP+F0	AR	/	RNN
EMPHASIS [195]	Ling→LinS+CAP+F0	AR	/	Hybrid
ARST [382]	Ph→LSP+BAP+F0	AR	Seq2Seq	RNN
VoiceLoop [339]	Ph→MGC+BAP+F0	AR	/	hybrid
RNN				
Tacotron [389]	Ch→LinS	AR	Seq2Seq	Hybrid/RNN
Tacotron 2 [309]	Ch→MelS	AR	Seq2Seq	RNN
DurIAN [426]	Ph→MelS	AR	Seq2Seq	RNN
Non-Att Tacotron [310]	Ph→MelS	AR	/	Hybrid/CNN/RNN
Para. Tacotron 1/2 [75, 76]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
MelNet [374]	Ch→MelS	AR	/	RNN
CNN				
DeepVoice [8]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 2 [88]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 3 [276]	Ch/Ph→MelS	AR	Seq2Seq	CNN
ParaNet [274]	Ph→MelS	NAR	Seq2Seq	CNN
DCTTS [338]	Ch→MelS	AR	Seq2Seq	CNN
SpeedySpeech [368]	Ph→MelS	NAR	/	CNN
TalkNet 1/2 [19, 18]	Ch→MelS	NAR	/	CNN
Transformer				
TransformerTTS [196]	Ph→MelS	AR	Seq2Seq	Self-Att
MultiSpeech [39]	Ph→MelS	AR	Seq2Seq	Self-Att
FastSpeech 1/2 [296, 298]	Ph→MelS	NAR	Seq2Seq	Self-Att
AlignTTS [437]	Ch/Ph→MelS	NAR	Seq2Seq	Self-Att
JDI-T [201]	Ph→MelS	NAR	Seq2Seq	Self-Att
FastPitch [185]	Ph→MelS	NAR	Seq2Seq	Self-Att
AdaSpeech 1/2/3 [40, 411, 412]	Ph→MelS	NAR	Seq2Seq	Self-Att
AdaSpeech 4 [399]	Ph→MelS	NAR	Seq2Seq	Self-Att
DenoiSpeech [442]	Ph→MelS	NAR	Seq2Seq	Self-Att
DeviceTTS [127]	Ph→MelS	NAR	/	Hybrid/DNN/RNN
LightSpeech [226]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
DelightfulTTS [216]	Ph→MelS	NAR	Seq2Seq	Self-Att
Flow				
Flow-TTS [240]	Ch/Ph→MelS	NAR*	Flow	Hybrid/CNN/RNN
Glow-TTS [162]	Ph→MelS	NAR	Flow	Hybrid/Self-Att/CNN
Flowtron [373]	Ph→MelS	AR	Flow	Hybrid/RNN
EfficientTTS [241]	Ch→MelS	NAR	Flow	Hybrid/CNN
VAE				
GMVAE-Tacotron [120]	Ph→MelS	AR	VAE	Hybrid/RNN
VAE-TTS [451]	Ph→MelS	AR	VAE	Hybrid/RNN
BVAE-TTS [191]	Ph→MelS	NAR	VAE	CNN
VARA-TTS [208]	Ph→MelS	NAR	VAE	CNN
GAN				
GAN exposure [100]	Ph→MelS	AR	GAN	Hybrid/RNN
TTS-Stylization [230]	Ch→MelS	AR	GAN	Hybrid/RNN
Multi-SpectroGAN [190]	Ph→MelS	NAR	GAN	Hybrid/Self-Att/CNN
Diffusion				
Diff-TTS [142]	Ph→MelS	NAR*	Diffusion	Hybrid/CNN
Grad-TTS [282]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN
PriorGrad [189]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN
Guided-TTS [161]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN
DiffGAN-TTS [215]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN

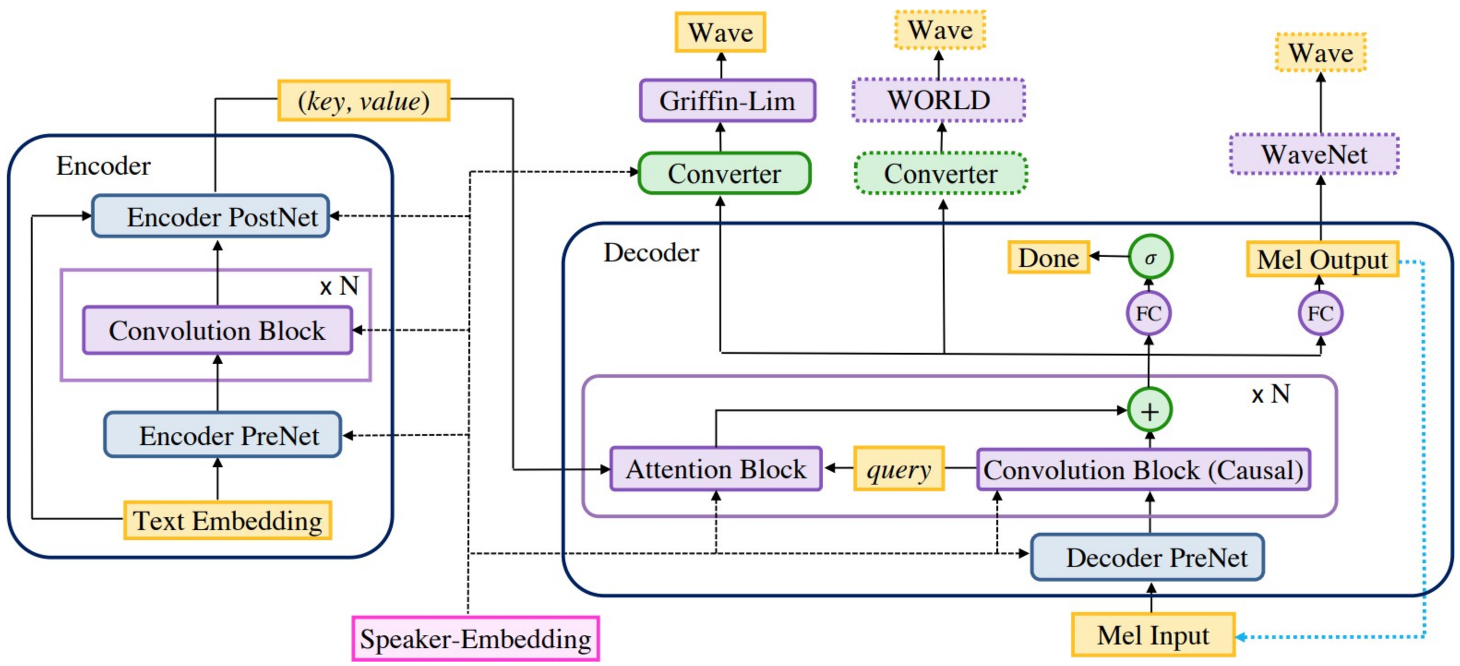
Acoustic model——RNN based

- Tacotron 2
  - Evolved from Tacotron
  - Text to mel-spectrogram generation
  - LSTM based encoder and decoder
  - Location sensitive attention
  - WaveNet as the vocoder



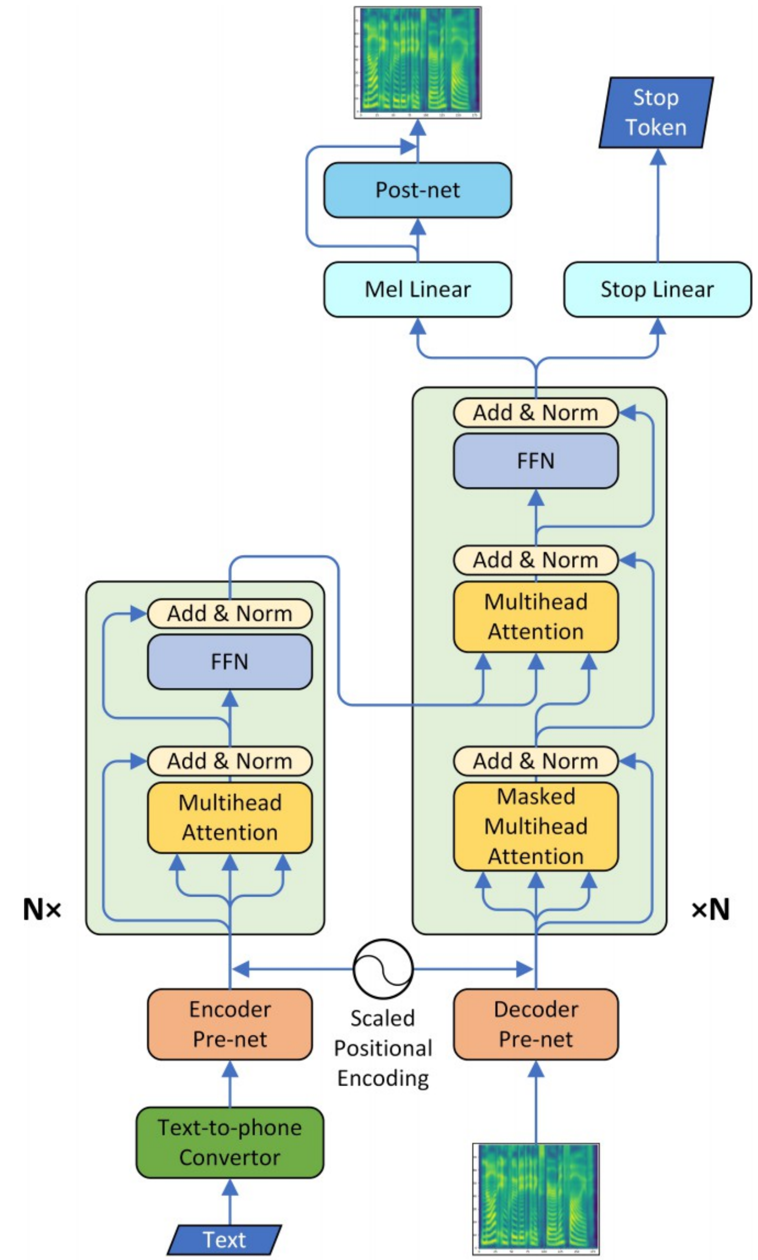
Acoustic model——CNN based

- DeepVoice 3
  - Evolved from DeepVoice 1/2
  - Enhanced with purely CNN based structure
  - Support different acoustic features as output
  - Support multi-speakers



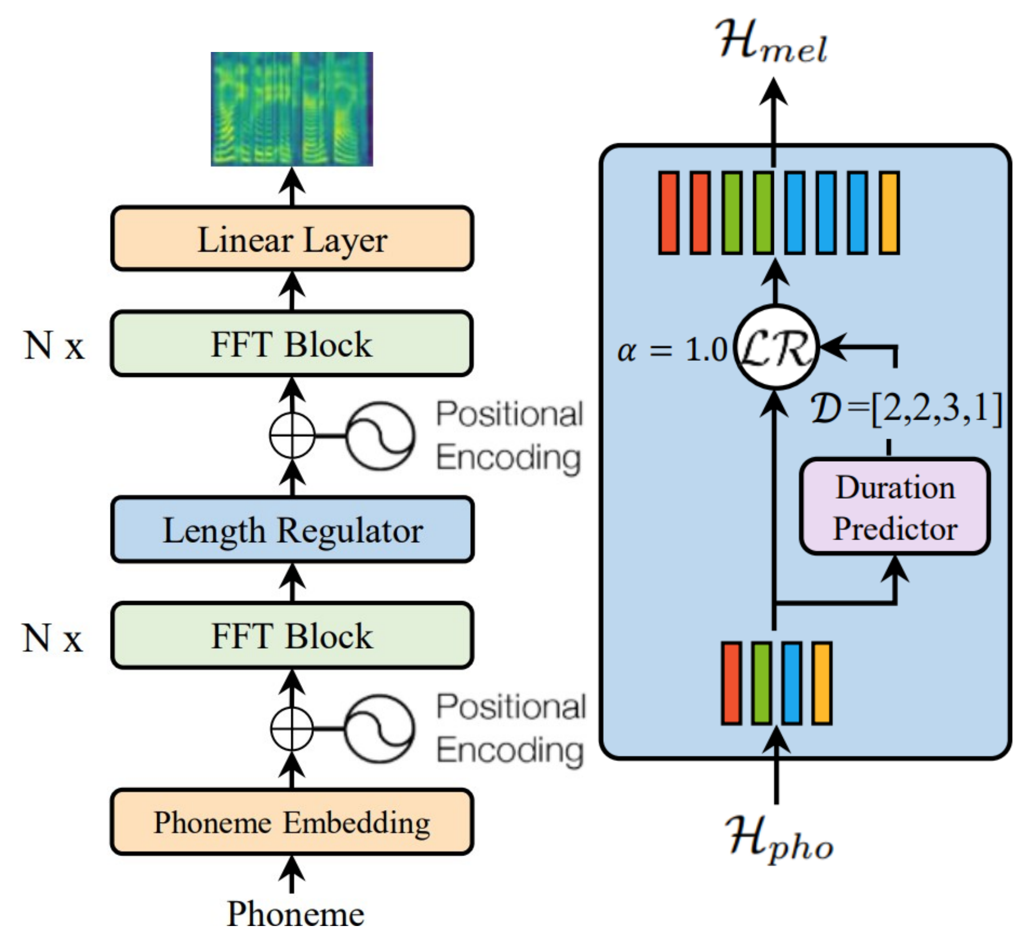
Acoustic model——Transformer based

- TransformerTTS
  - Framework is like Tacotron 2
  - Replace LSTM with Transformer in encoder and decoder
  - Parallel training, quality on par with Tacotron 2
  - Attention with more challenges than Tacotron 2, due to parallel computing



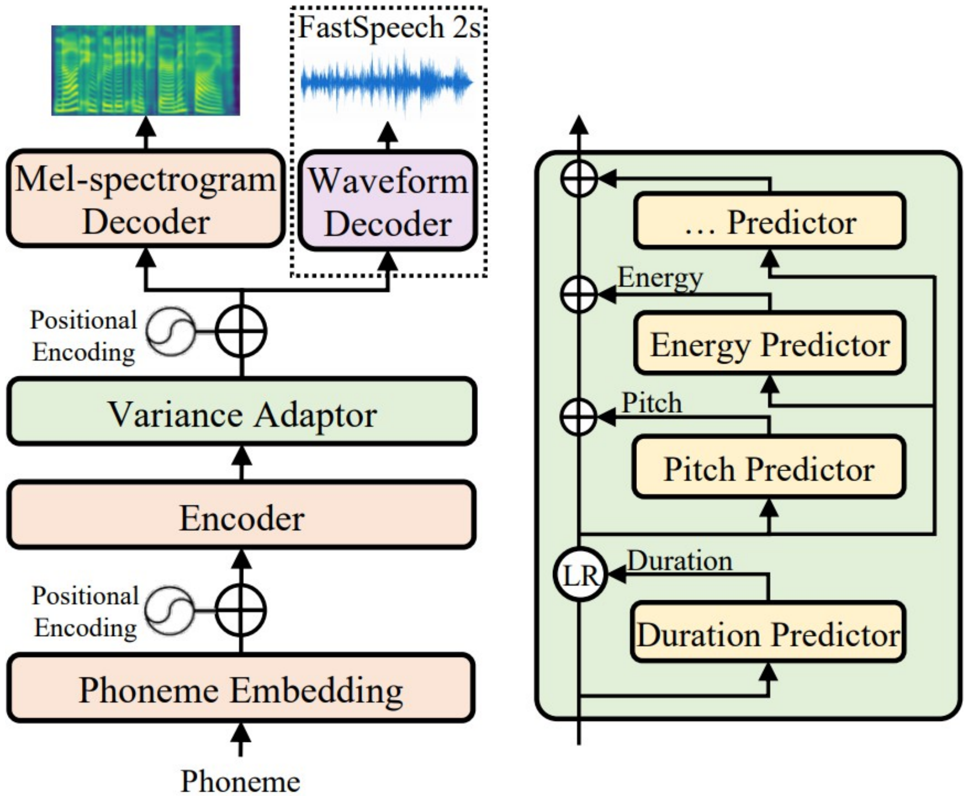
Acoustic model——Transformer based

- FastSpeech [290]
  - Generate mel-spectrogram in parallel (for speedup)
  - Remove the text-speech attention mechanism (for robustness)
  - Feed-forward transformer with length regulator (for controllability)



Acoustic model——Transformer based

- FastSpeech 2 [292]
  - Improve FastSpeech
  - Use variance adaptor to predict duration, pitch, energy, etc
  - Simplify training pipeline of FastSpeech (KD)
  - FastSpeech 2s: a fully end-to-end parallel text to wave model



(a) FastSpeech 2 (b) Variance adaptor

- Other works
  - FastPitch [181]
  - JDI-T [197], AlignTTS [429]

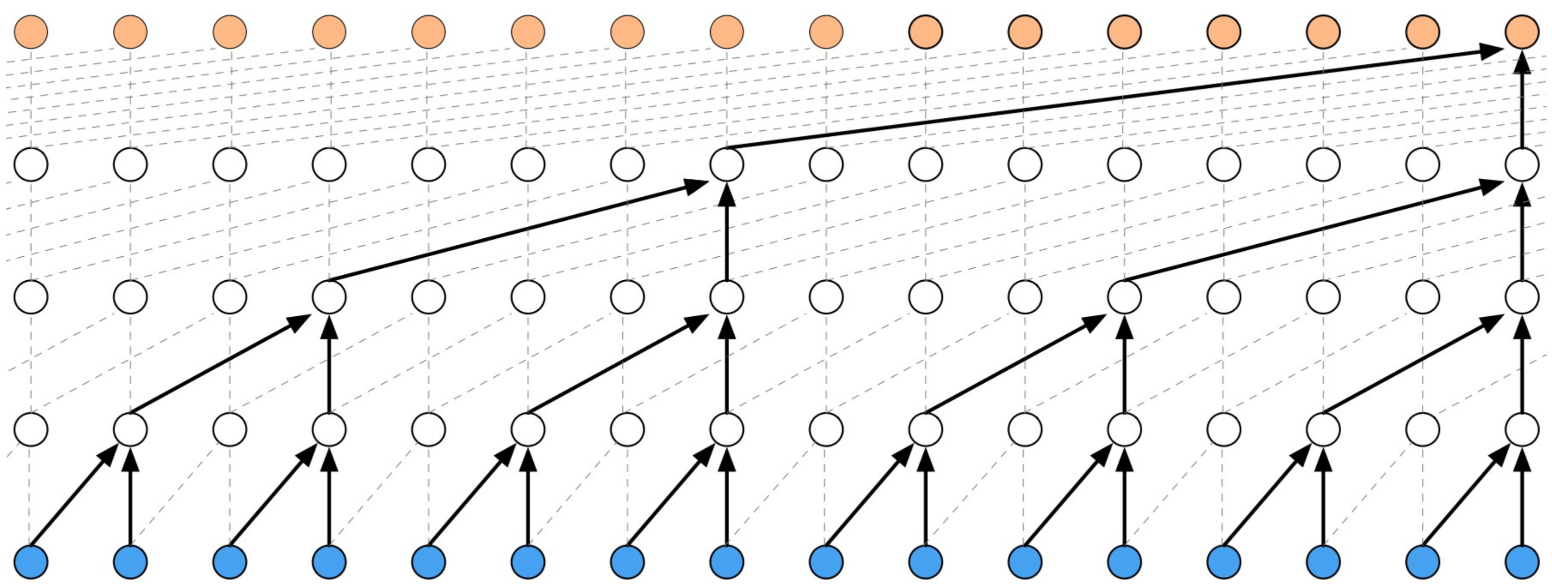
Vocoder

- Autoregressive vocoder
- Flow-based vocoder
- GAN-based vocoder
- VAE-based vocoder
- Diffusion-based vocoder

	Vocoder	Input	AR/NAR	Modeling	Architecture
AR	WaveNet [260]	Linguistic Feature	AR	/	CNN
	SampleRNN [239]	/	AR	/	RNN
	WaveRNN [151]	Linguistic Feature	AR	/	RNN
	LPCNet [370]	BFCC	AR	/	RNN
	Univ. WaveRNN [221]	Mel-Spectrogram	AR	/	RNN
	SC-WaveRNN [271]	Mel-Spectrogram	AR	/	RNN
	MB WaveRNN [426]	Mel-Spectrogram	AR	/	RNN
	FFTNet [146]	Cepstrum	AR	/	CNN
	iSTFTNet [153]	Mel-Spectrogram	NAR	/	CNN
Flow	Par. WaveNet [261]	Linguistic Feature	NAR	Flow	CNN
	WaveGlow [285]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
	FloWaveNet [166]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
	WaveFlow [277]	Mel-Spectrogram	AR	Flow	Hybrid/CNN
	SqueezeWave [441]	Mel-Spectrogram	NAR	Flow	CNN
GAN	WaveGAN [69]	/	NAR	GAN	CNN
	GELP [150]	Mel-Spectrogram	NAR	GAN	CNN
	GAN-TTS [23]	Linguistic Feature	NAR	GAN	CNN
	MelGAN [182]	Mel-Spectrogram	NAR	GAN	CNN
	Par. WaveGAN [410]	Mel-Spectrogram	NAR	GAN	CNN
	HiFi-GAN [178]	Mel-Spectrogram	NAR	GAN	Hybrid/CNN
	VocGAN [416]	Mel-Spectrogram	NAR	GAN	CNN
	GED [97]	Linguistic Feature	NAR	GAN	CNN
	Fre-GAN [164]	Mel-Spectrogram	NAR	GAN	CNN
VAE	Wave-VAE [274]	Mel-Spectrogram	NAR	VAE	CNN
Diffusion	WaveGrad [41]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
	DiffWave [180]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
	PriorGrad [189]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
	SpecGrad [176]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN

Vocoder—AR

- WaveNet: autoregressive model with dilated causal convolution [van den oord et al 2016]



- Other works
  - WaveRNN
  - LPCNet

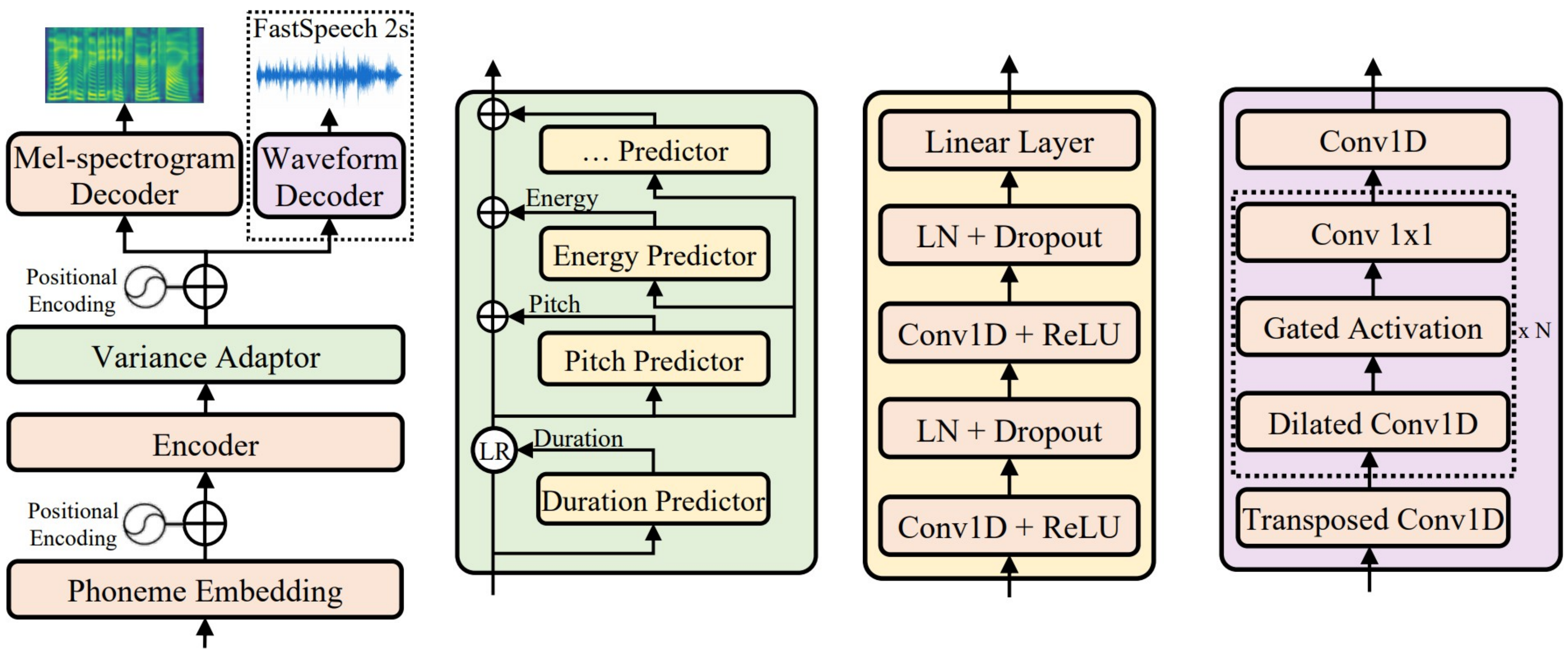
Generative models for acoustic model/vocoder

- Text to speech mapping p(x|y) is multimodal, since one text can correspond to multiple speech variations
  - Acoustic model, phoneme-spectrogram mapping: duration/pitch/energy/formant
  - Vocoder, spectrogram-waveform mapping: phase

- How to model a multimodal conditional distribution p(x|y)?
  - Autoregressive, GAN, VAE, Flow, Diffusion Model, etc
  - Since L1/L2 can be applied to mel-spectrogram, while cannot be directly applied to waveform
  - Advanced generative models are developed faster in vocoder than in acoustic model, but finally acoustic models catch up ☺

Fully End-to-End TTS

- FastSpeech 2s: fully parallel text to wave model



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor (d) Waveform decoder

Current developments in TTS

- Fully Controllable TTS
  - Identity
  - Dialect
  - Contextual Appropriateness

- Explainable, light weight models
  - Grounded in Human mechanisms
  - Edge Deployable
  - Seamless interfacing with other Modalities