REVIEW OF TEXT-TO-SPEECH CONVERSION FOR ENGLISH

Presented by
Najeeb Khan
2013-3-8

INTRODUCTION

INTRODUCTION

Trace the history of progress toward the development of systems for converting text to speech.

TEXT TO SPEECH

TEXT TO SPEECH

TTS involves two steps

TEXT TO SPEECH

TTS involves two steps

- A set of modules must analyze the text to determine the underlying structure of the sentence and the phonemic composition of each word.

TEXT TO SPEECH

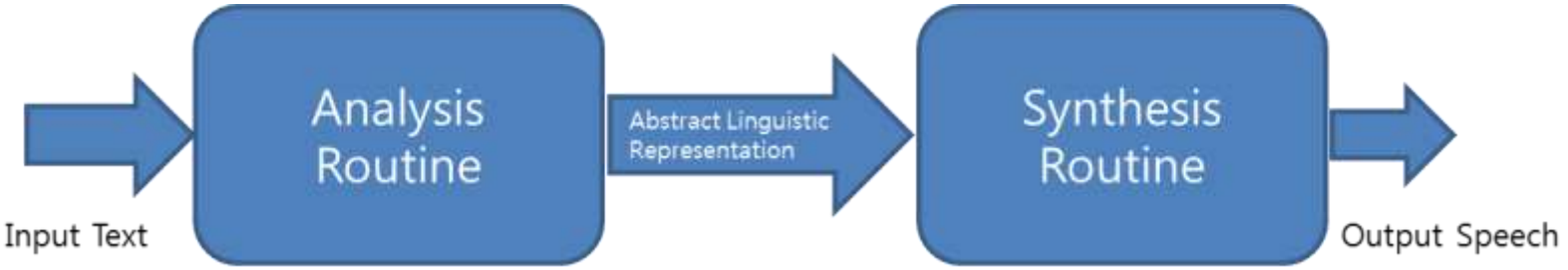
TTS involves two steps

- A set of modules must analyze the text to determine the underlying structure of the sentence and the phonemic composition of each word.
- A second set of modules that transforms this abstract linguistic representation into speech waveform.

TEXT TO SPEECH

TTS involves two steps

- A set of modules must analyze the text to determine the underlying structure of the sentence and the phonemic composition of each word.
- A second set of modules that transforms this abstract linguistic representation into speech waveform.



A SIMPLE APPROACH TO TTS

A SIMPLE APPROACH TO TTS

- Store natural waveforms corresponding to each word

A SIMPLE APPROACH TO TTS

- Store natural waveforms corresponding to each word
- Simply concatenate them to produce sentences

DRAWBACKS

DRAWBACKS

- Words are as short as half their duration when spoken in isolation, making concatenated speech painfully slow.

DRAWBACKS

- Words are as short as half their duration when spoken in isolation, making concatenated speech painfully slow.
- Stress pattern, Rhythm, and Intonation are very unnatural

DRAWBACKS

- Words are as short as half their duration when spoken in isolation, making concatenated speech painfully slow.
- Stress pattern, Rhythm, and Intonation are very unnatural
- Words blend together at an articulatory level

DRAWBACKS

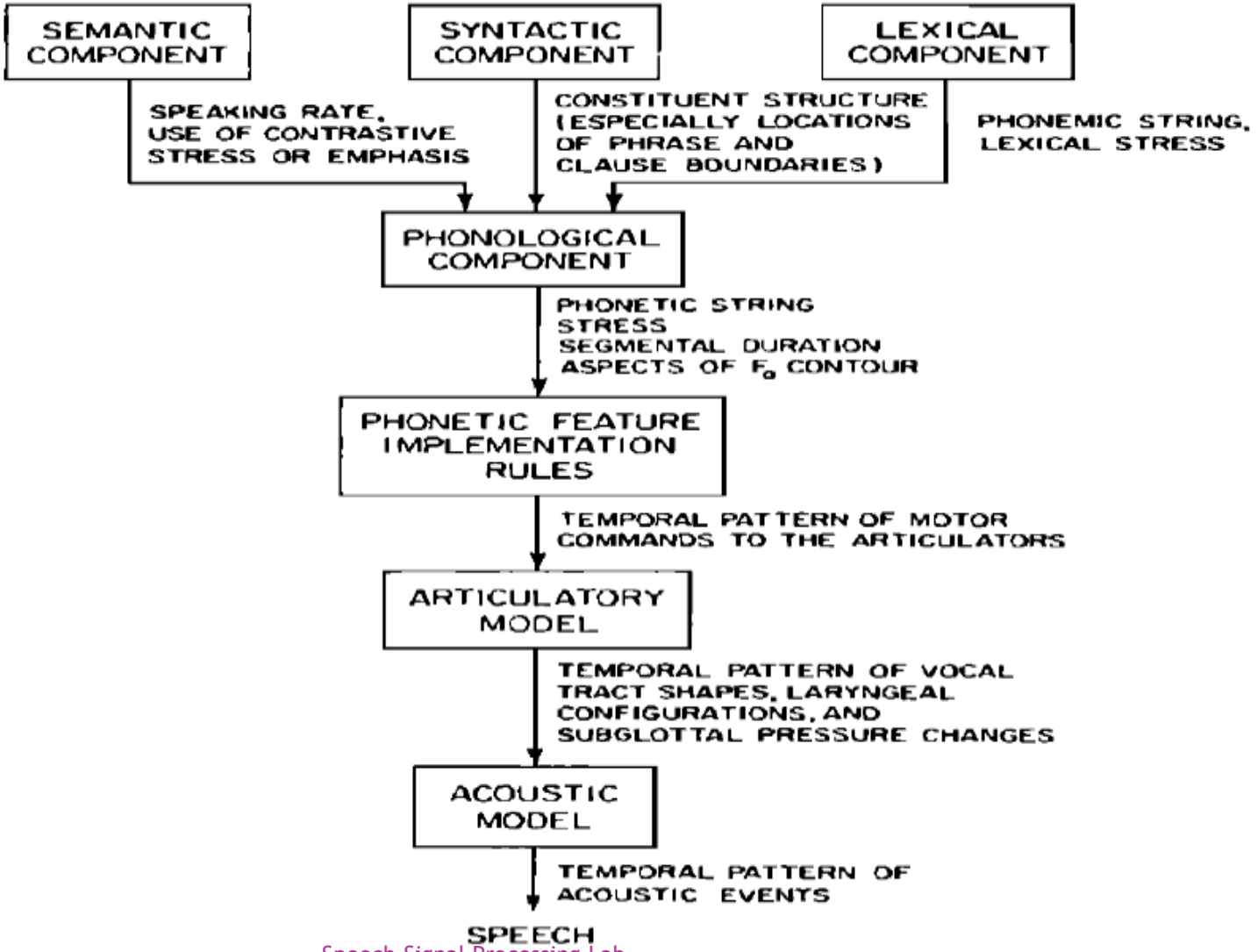
- Words are as short as half their duration when spoken in isolation, making concatenated speech painfully slow.
- Stress pattern, Rhythm, and Intonation are very unnatural
- Words blend together at an articulatory level
- Number of words is extremely large and new words are coined everyday

LINGUISTIC FRAMEWORK

LINGUISTIC FRAMEWORK

- Generative linguists specify rules for the generation of any legitimate sentence of the language.

LINGUISTIC FRAMEWORK



CONTD...

CONTD…

- A sentence can be represented by a sequence of discrete elements called Phonemes.

CONTD…

- A sentence can be represented by a sequence of discrete elements called Phonemes.
- Phonemes strings form larger units such as syllables, words, phrases and clauses.

CONTD…

- A sentence can be represented by a sequence of discrete elements called Phonemes.
- Phonemes strings form larger units such as syllables, words, phrases and clauses.
- Boundaries of these structures helps in the pronunciation

CONTD…

- A sentence can be represented by a sequence of discrete elements called Phonemes.
- Phonemes strings form larger units such as syllables, words, phrases and clauses.
- Boundaries of these structures helps in the pronunciation
- Each syllable of a word in a sentence can be assigned a stress level

CONTD...

- A sentence can be represented by a sequence of discrete elements called Phonemes.
- Phonemes strings form larger units such as syllables, words, phrases and clauses.
- Boundaries of these structures helps in the pronunciation
- Each syllable of a word in a sentence can be assigned a stress level
- The stress pattern changes the durations of sounds and pitch over an utterance.

CONTD...

CONTD…

- The phonological component converts information about stress and boundary types into:

CONTD…

- The phonological component converts information about stress and boundary types into:
  - A string of phonetic segments

CONTD…

- The phonological component converts information about stress and boundary types into:
  - A string of phonetic segments
  - A superimposed pattern of timing, intensity and f_0 motions.

PHONEMES TO SPEECH CONVERSION

PHONEMES TO SPEECH CONVERSION

Example: “Joe ate his soup”

ABSTRACT LINGUISTIC REPRESENTATION:

/ǰo) ˈet hɪz sˈup./

ALLOPHONIC RECODING:

[ǰo ˈeɾ ɛs sˈup.]

DURATION SPECIFICATION, IN MSEC:

[100, 210, 180, 20, 65, 75, 90, 165, 75]

FUNDAMENTAL FREQUENCY GESTURES:

1. HAT RISE DURING [o]
2. STRESS PULSE ON [o]
3. STRESS PULSE ON [e]
4. STRESS PULSE ON [u]
5. HAT FALL DURING [u]

CONTD...

CONTD...

- Each phonetic segment is assigned an inherent duration by table lookup. A set of rules is applied to predict changes to the duration of the segment as a function of sentential context.

CONTD…

- Each phonetic segment is assigned an inherent duration by table lookup. A set of rules is applied to predict changes to the duration of the segment as a function of sentential context.
- A fundamental frequency contour is determined by rules that specify the locations and amplitudes of step and impulse commands.

CONTD...

CONTD…

- A phonetic synthesis by rule system derives time functions that characterize the activity of voicing and noise sources and the acoustic resonance properties of the vocal tract.(19 time functions in Klattalk)

CONTD…

- A phonetic synthesis by rule system derives time functions that characterize the activity of voicing and noise sources and the acoustic resonance properties of the vocal tract.(19 time functions in Klattalk)
- Finally a formant synthesizer is used to convert this parametric representation into a speech waveform.

EARLY SYNTHESIZER

EARLY SYNTHESIZER

- Stewart 1922: Two resonant circuits were excited by a buzzer permitting approximations to static vowel sound.

EARLY SYNTHESIZER

- Stewart 1922: Two resonant circuits were excited by a buzzer permitting approximations to static vowel sound.

- Vocoder/Voder: A device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct an approximation to the orignal waveform.

PATTERN PLAYBACK SYNTHESIZER

PATTERN PLAYBACK SYNTHESIZER

- A wheel generates harmonics of 120Hz tone, while harmonic amplitudes are controlled over time by the reflectance of painted spectrographic patterns on a moving transparent belt.

SOURCE FILTER THEORY OF SPEECH GENERATION

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.
- Sources:

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.
- Sources:
  - Vibration of vocal folds

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.
- Sources:
  - Vibration of vocal folds
  - Turbulence noise

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.
- Sources:
  - Vibration of vocal folds
  - Turbulence noise
- Linear Filter:

SOURCE FILTER THEORY OF SPEECH GENERATION

- It is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources.
- Sources:
  - Vibration of vocal folds
  - Turbulence noise
- Linear Filter:
  - Simulates the resonance effects of the acoustic tube formed by pharynx, oral cavity and lips. Poles creates local peaks called formants.

PARALLEL FORMANT SYNTHESIZER

PARALLEL FORMANT SYNTHESIZER

- The outputs of a set of resonators connected in parallel are summed and the input sound source amplitude of each formant resonator is determined by an independent control parameter.

PARAMETRIC ARTIFICIAL TALKER

PARAMETRIC ARTIFICIAL TALKER

- PAT consisted of three electronic formant resonators connected in parallel, whose inputs were either a buzz or noise.

PARAMETRIC ARTIFICIAL TALKER

- PAT consisted of three electronic formant resonators connected in parallel, whose inputs were either a buzz or noise.

- A moving glass slide was used to convert painted patterns into 3 formants, A_v , f_0 and A_n.

PARAMETRIC ARTIFICIAL TALKER

- PAT consisted of three electronic formant resonators connected in parallel, whose inputs were either a buzz or noise.
- A moving glass slide was used to convert painted patterns into 3 formants, A_v , f_0 and A_n.
- PAT was latter modified to have a separate circuit for fricatives and converted to cascade operation.

ORATOR VERBIS ELECTRIS

ORATOR VERBIS ELECTRIS

- OVE I consisted of formant resonators connected in series, the lowest two of which were varied in frequency by movements in two dimensions of a mechanical arm. The amplitude and f_0 were manually controlled.

OVE II

OVE II

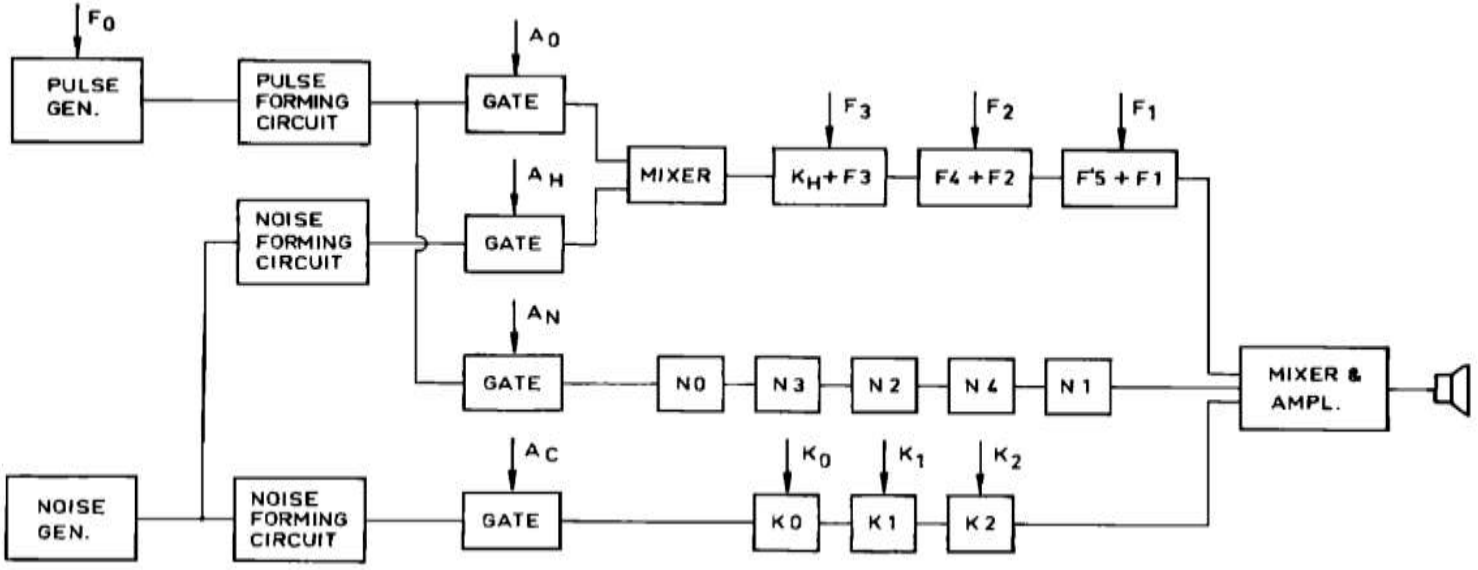
- OVE II included a separate static branch to simulate nasals.

OVE II

- OVE II included a separate static branch to simulate nasals.
- Cascade of two formants and one anti-formant to simulate a simplified approximation to the vocal tract.

OVE II

- OVE II included a separate static branch to simulate nasals.
- Cascade of two formants and one anti-formant to simulate a simplified approximation to the vocal tract.



KLATT SYNTHESIZER

KLATT SYNTHESIZER

- Klatt synthesizer include

KLATT SYNTHESIZER

- Klatt synthesizer include
  - Cascaded formants for synthesis of sonorants

KLATT SYNTHESIZER

- Klatt synthesizer include
  - Cascaded formants for synthesis of sonorants
  - Parallel formants for the synthesis of obstruents

KLATT SYNTHESIZER

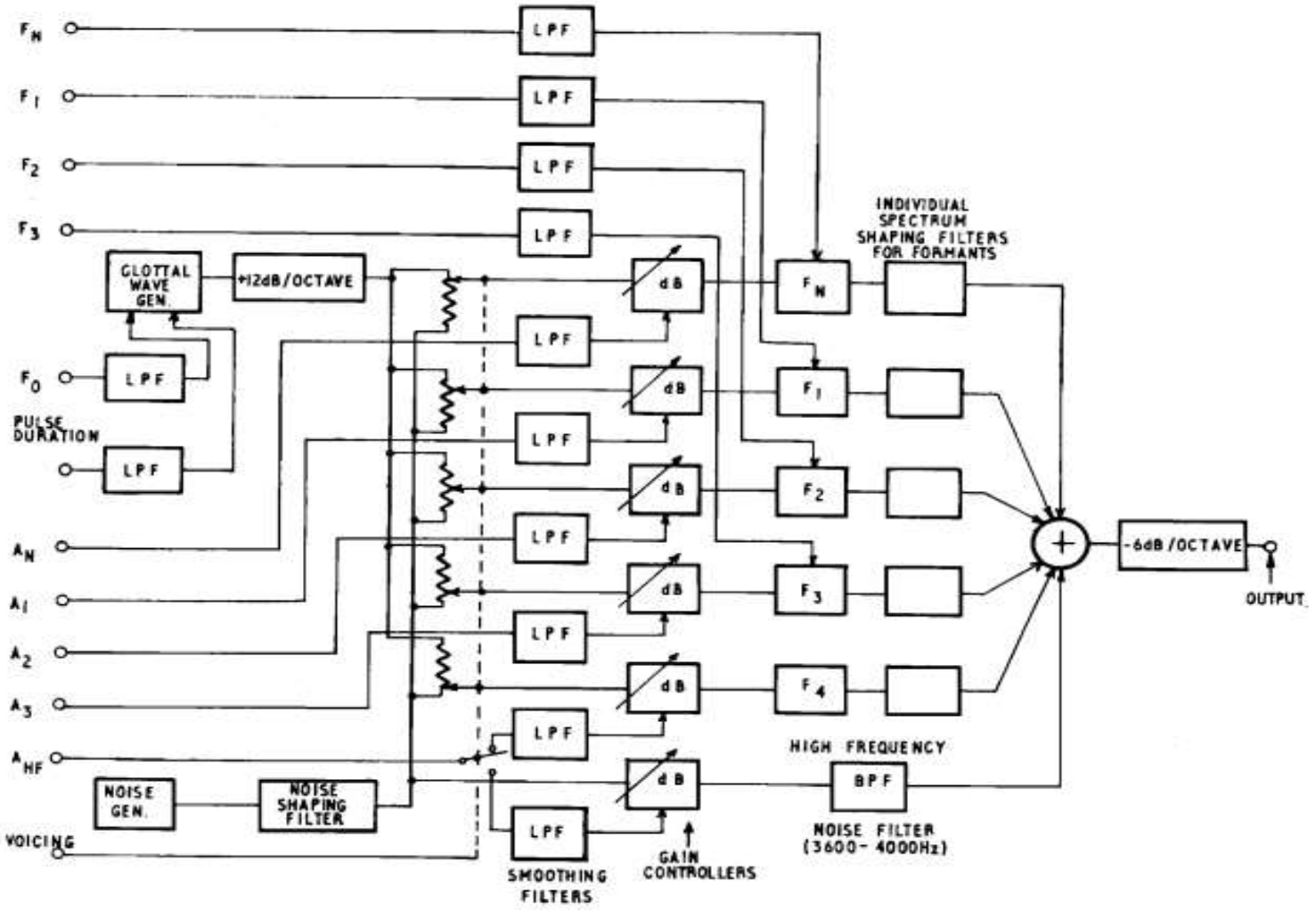
- Klatt synthesizer include
  - Cascaded formants for synthesis of sonorants
  - Parallel formants for the synthesis of obstruents
  - An extra pole-zero for nasalization

HOLMES SYNTHESIZER

HOLMES SYNTHESIZER

- The objective of Holmes synthesizer is to match a natural recording recording of a particular speaker

HOLMES SYNTHESIZER



CONTD...

CONTD...

- It is desirable to use a voicing waveform based on that of the speaker being modeled.

CONTD…

- It is desirable to use a voicing waveform based on that of the speaker being modeled.
- Holmes found that variability in the spectra of natural speech can be mimicked by proper adjustments to the amplitudes of parallel formants.

CONTD...

- It is desirable to use a voicing waveform based on that of the speaker being modeled.
- Holmes found that variability in the spectra of natural speech can be mimicked by proper adjustments to the amplitudes of parallel formants.
- Irregularities in spectrum between formant peaks are of little perceptual importance, only strong harmonics near a formant peak and below f_1 must be synthesized.

MODELS OF THE VOICING SOURCE

MODELS OF THE VOICING SOURCE

- Early Voicing Sources:

MODELS OF THE VOICING SOURCE

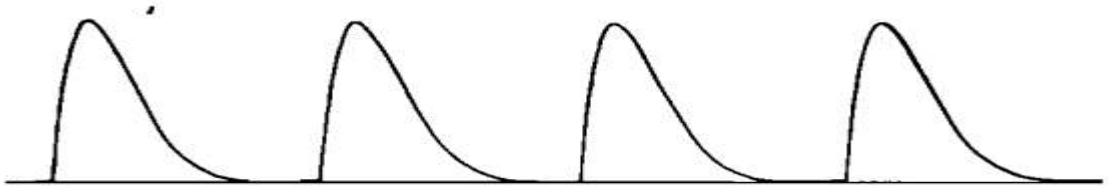
- Early Voicing Sources:
  - Sawtooth waveforms

MODELS OF THE VOICING SOURCE

- Early Voicing Sources:
  - Sawtooth waveforms
  - Filtered impulse train

MODELS OF THE VOICING SOURCE

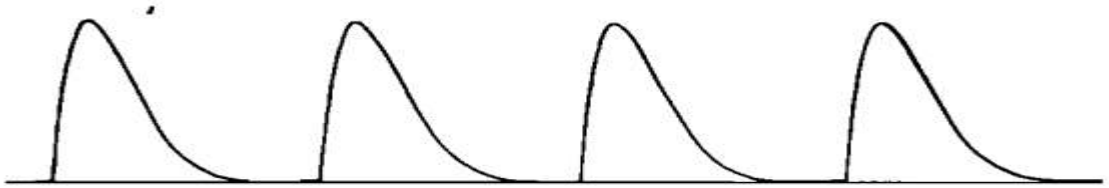
- Early Voicing Sources:
  - Sawtooth waveforms
  - Filtered impulse train



LOW-PASS FILTERED IMPULSE TRAIN

MODELS OF THE VOICING SOURCE

- Early Voicing Sources:
  - Sawtooth waveforms
  - Filtered impulse train



LOW-PASS FILTERED IMPULSE TRAIN

- Spectrum of these waveforms is right but the phase is wrong.

MODELS OF THE VOICING SOURCE

- Early Voicing Sources:
  - Sawtooth waveforms
  - Filtered impulse train

LOW-PASS FILTERED IMPULSE TRAIN

- Spectrum of these waveforms is right but the phase is wrong.
- Spectrum is monotonic, contrasting the presence of zeros in the spectrum of normal voicing waveforms.

CONTD...

CONTD...

- Rothenberg et al (1975):

CONTD...

- Rothenberg et al (1975):
  - 3 parameter model producing waveshapes varying w.r.t f_0, Amplitude, OQ, degree of static glottal opening and breathness.

CONTD...

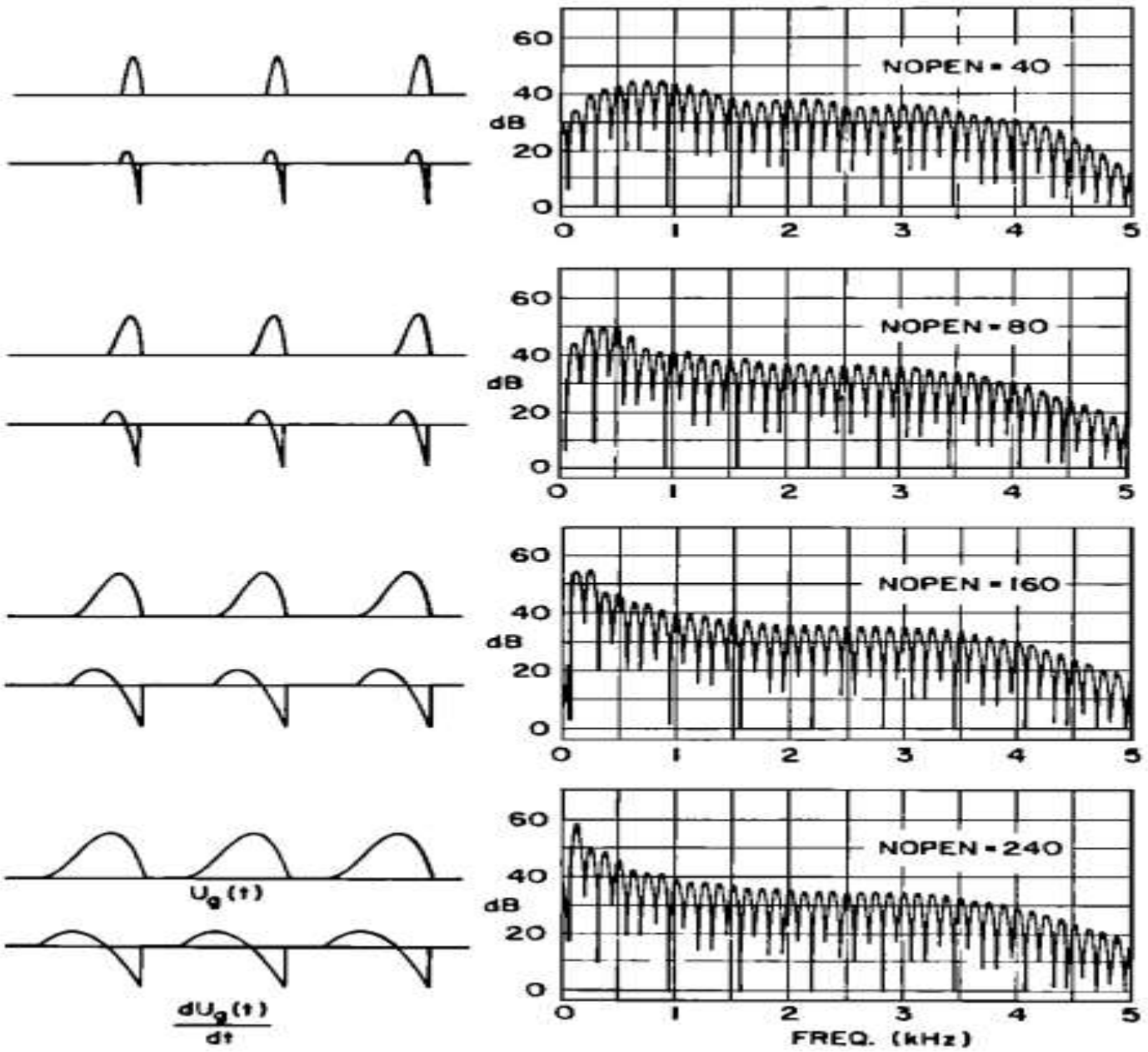
- Rothenberg et al (1975):
  - 3 parameter model producing waveshapes varying w.r.t f_0, Amplitude, OQ, degree of static glottal opening and breathness.

- Fant et al (1985):

CONTD…

- Rothenberg et al (1975):
  - 3 parameter model producing waveshapes varying w.r.t f_0, Amplitude, OQ, degree of static glottal opening and breathness.

- Fant et al (1985):
  - Similar to Rothenberg model but with more control over the important acoustic variables (spectral tilt, zeros location & intensity of f_0).


NOPEN = 40
NOPEN = 80
NOPEN = 160
NOPEN = 240
dB
60
40
20
0
0
1
2
3
4
5
FREQ. (kHz)
$U_g(t)$
$\frac{dU_g(t)}{dt}$

CONTD...

CONTD...

- Klattalk Model: Glottal waveform can be modified in

CONTD…

- Klattalk Model: Glottal waveform can be modified in
  - Open period

CONTD…

- Klattalk Model: Glottal waveform can be modified in
  - Open period
  - Abruptness of the closing

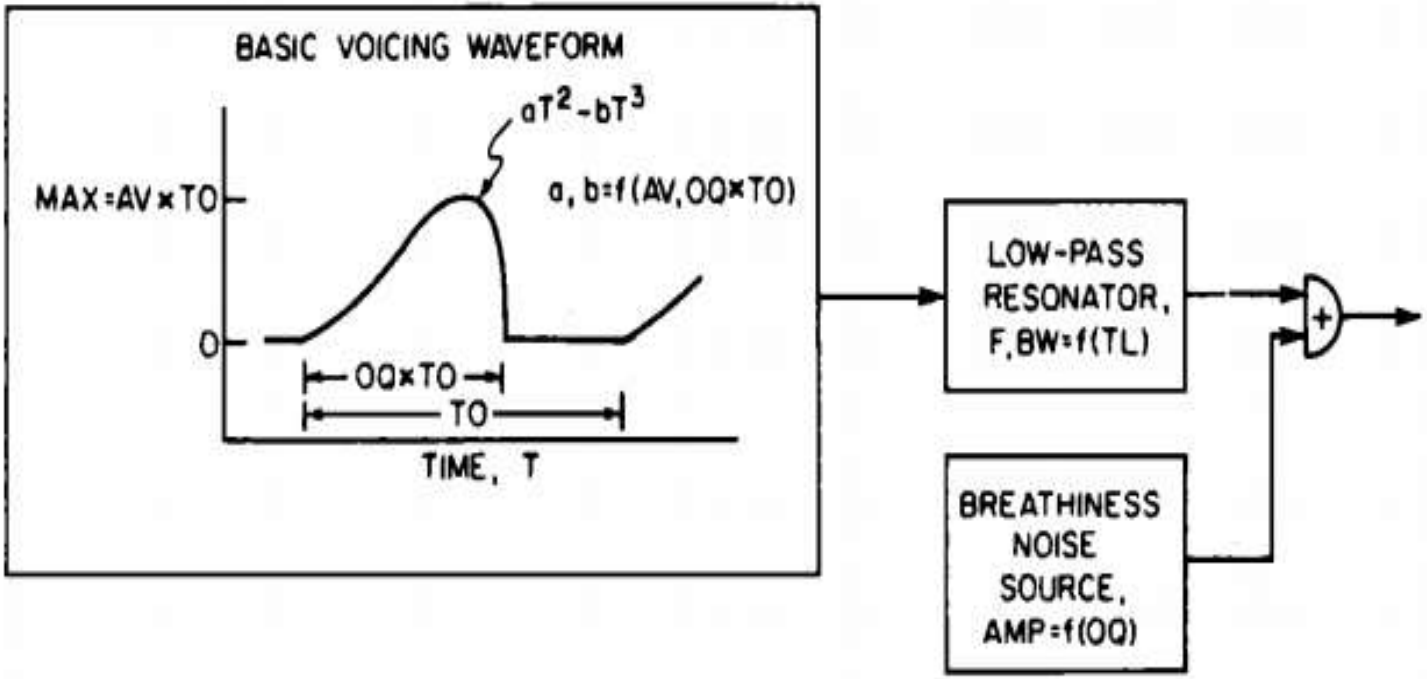
CONTD…

- Klattalk Model: Glottal waveform can be modified in
  - Open period
  - Abruptness of the closing
  - Breathness

CONTD…

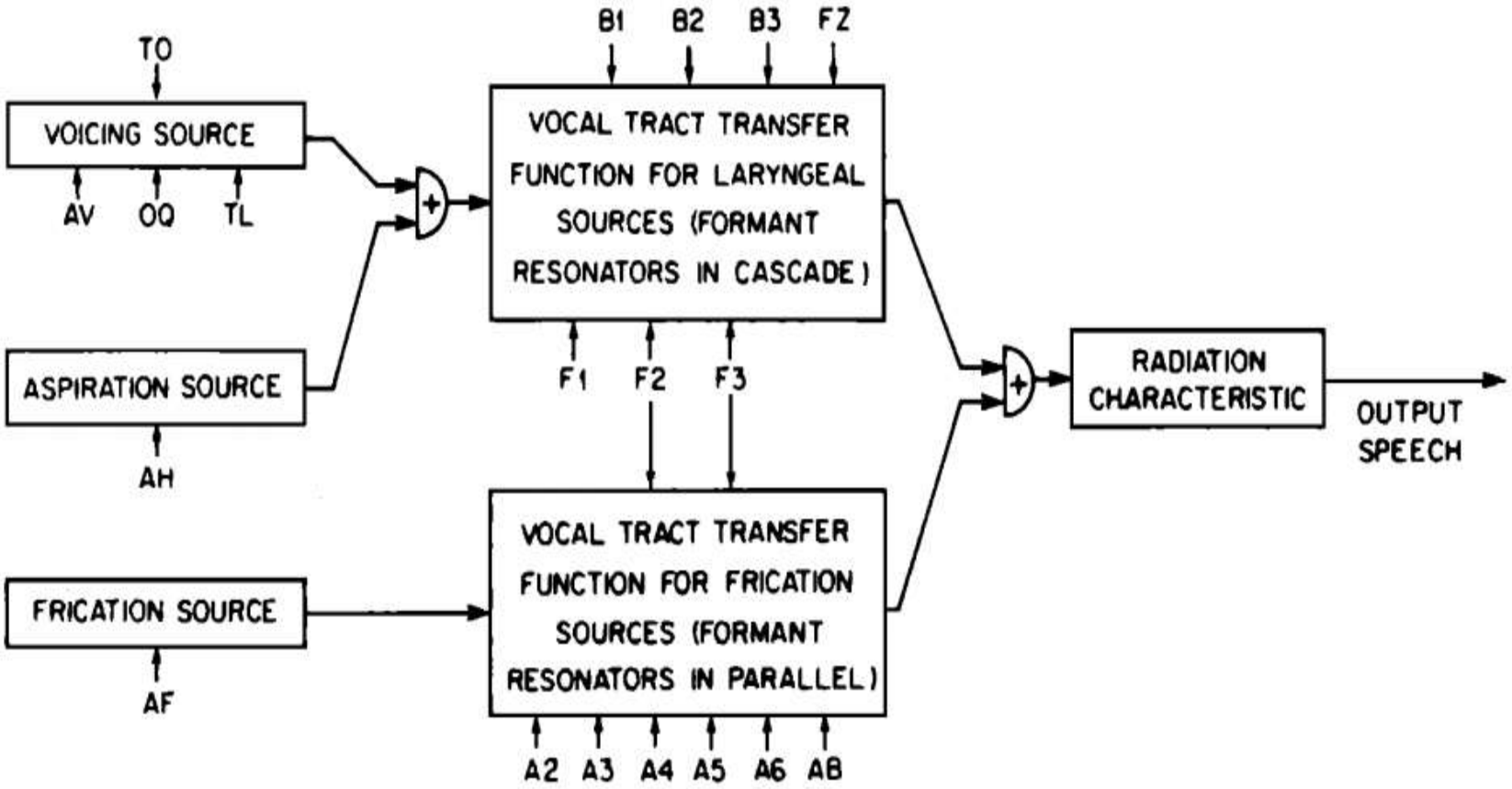
- Klattalk Model: Glottal waveform can be modified in
  - Open period
  - Abruptness of the closing
  - Breathness
  - Degree of diplo-phonic vibration

CONTD...


BASIC VOICING WAVEFORM
aT^2-bT^3
a, b=f(AV, OQ×TO)
MAX=AV×TO
0
OQ×TO
TO
TIME, T
LOW-PASS RESONATOR, F, BW=f(TL)
BREATHINESS NOISE SOURCE, AMP=f(OQ)
+

KLATTALK SYNTHESIZER

KLATTALK SYNTHESIZER



KLATTALK SYNTHESIZER

KLATTALK SYNTHESIZER

- Limited Naturalness

KLATTALK SYNTHESIZER

- Limited Naturalness
  - Either the model is not correct

KLATTALK SYNTHESIZER

- Limited Naturalness
  - Either the model is not correct
  - Or the control mechanism is not proper

FLANAGAN WORK

FLANAGAN WORK

- Described the expected locations of voicing source zeros as a function of various assumptions about the nature of the glottal volume velocity waveform.

FLANAGAN WORK

- Described the expected locations of voicing source zeros as a function of various assumptions about the nature of the glottal volume velocity waveform.
- Flanagan shows that the frequency locations and depths of spectral notches induced by source zeros depend on relatively small changes to critical aspects of the source waveform such as symmetry.

FLANAGAN WORK

- Described the expected locations of voicing source zeros as a function of various assumptions about the nature of the glottal volume velocity waveform.
- Flanagan shows that the frequency locations and depths of spectral notches induced by source zeros depend on relatively small changes to critical aspects of the source waveform such as symmetry.
- Holmes didn't followed these details but followed changes observed in natural speech.

FEMALE VOICE SYNTHESIS

FEMALE VOICE SYNTHESIS

- Simple scaling procedures do not result in a particular female voice quality, non-uniform formant scaling appears to be required.

FEMALE VOICE SYNTHESIS

- Simple scaling procedures do not result in a particular female voice quality, non-uniform formant scaling appears to be required.
- Analysis of female speech revealed the presence of considerable random breathiness noise above 2KHz & considerable variation in tilt and f_0 magnitude.

FEMALE VOICE SYNTHESIS

- Simple scaling procedures do not result in a particular female voice quality, non-uniform formant scaling appears to be required.
- Analysis of female speech revealed the presence of considerable random breathiness noise above 2KHz & considerable variation in tilt and f_0 magnitude.
- Klattalk model achieved good approximation to female voice for vowels.

CONTD...

CONTD...

- For complex speech, utterances involving a glottal stop were easier to model.

CONTD…

- For complex speech, utterances involving a glottal stop were easier to model.
- For breathy vowels [hv], many of the voiced intervals revealed additional formant peaks and other harmonic amplitude discrepancies.

CONTD…

- For complex speech, utterances involving a glottal stop were easier to model.
- For breathy vowels [hv], many of the voiced intervals revealed additional formant peaks and other harmonic amplitude discrepancies.
- Presumably related to acoustic coupling with the tracheal resonances when the glottis is partially open.

CONTD...

CONTD…

- How best to augment the synthesizer in order to model the sudden appearance of additional formants and zeros in breathy vowels?

CONTD...

- How best to augment the synthesizer in order to model the sudden appearance of additional formants and zeros in breathy vowels?
- Synthesizer augmented with extra tracheal pole-zero pair has met with some success..

CONTD…

- How best to augment the synthesizer in order to model the sudden appearance of additional formants and zeros in breathy vowels?
- Synthesizer augmented with extra tracheal pole-zero pair has met with some success..
- Alternatively employ articulatory model of the trachea, vocal folds and vocal tract as well as their interactions in an articulatory synthesizer.