FROM TEXT TO SPEECH
THE MITALK SYSTEM

Presented by
Najeeb Khan
2013-6-7

THE FUNDAMENTAL FREQUENCY GENERATOR

THE FUNDAMENTAL FREQUENCY GENERATOR

- Attributes such as syntactic structure, emphasis, and sentence type can be partially signaled by the fundamental frequency contour

THE FUNDAMENTAL FREQUENCY GENERATOR

- Attributes such as syntactic structure, emphasis, and sentence type can be partially signaled by the fundamental frequency contour
- The f0 contour generator utilizes the phrase structure of each sentence as analyzed by the parser to determine declination lines, to calculate the amount of deviation from the declination line through each phrase and to insert continuation rises

THE FUNDAMENTAL FREQUENCY GENERATOR

THE FUNDAMENTAL FREQUENCY GENERATOR

- Lexical stress marks and syllable division are used to determine the location of f0 peaks

THE FUNDAMENTAL FREQUENCY GENERATOR

- Lexical stress marks and syllable division are used to determine the location of f0 peaks
- POS provide information needed to determine the relative height of peaks

THE FUNDAMENTAL FREQUENCY GENERATOR

- Lexical stress marks and syllable division are used to determine the location of f0 peaks
- POS provide information needed to determine the relative height of peaks
- Phonemic data provide information needed to determine segmental influences on f0

THE FUNDAMENTAL FREQUENCY GENERATOR

THE FUNDAMENTAL FREQUENCY GENERATOR

- The current algorithm produces two f0 target values for each phonetic segment, one to be used at onset and one as a mid-value

THE FUNDAMENTAL FREQUENCY GENERATOR

- The current algorithm produces two f0 target values for each phonetic segment, one to be used at onset and one as a mid-value
- The rises and falls which are calculated for each segment are used to specify target values, the peak point at either the left or right boundary of stressed vowels in content words and the midpoint target value for other segments

THE FUNDAMENTAL FREQUENCY GENERATOR

THE FUNDAMENTAL FREQUENCY GENERATOR

- Input

THE FUNDAMENTAL FREQUENCY GENERATOR

- Input
- The output file from the parser provides phrase group information and POS of individual words

THE FUNDAMENTAL FREQUENCY GENERATOR

- Input
- The output file from the parser provides phrase group information and POS of individual words
- Phrase groups which are recognized are noun phrases, prepositional phrases, verb phrases, and verbal groups

THE FUNDAMENTAL FREQUENCY GENERATOR

- Input
- The output file from the parser provides phrase group information and POS of individual words
- Phrase groups which are recognized are noun phrases, prepositional phrases, verb phrases, and verbal groups
- The POS of speech are grouped so as to be more useful in determining how they affect the f0 contour

THE FUNDAMENTAL FREQUENCY GENERATOR

Level	Part of speech
0	article
1	conjunction, relative pronoun
2	preposition, auxiliary verb, (unstressable modal, vocative)
3	personal pronoun
6	verb, demonstrative pronoun
7	noun, adjective, adverb, contraction
8	(reflexive pronoun)
9	stressable modal
10	quantifier
11	interrogative adjectives
12	(negative element)
14	(sentential adverb)

THE FUNDAMENTAL FREQUENCY GENERATOR

THE FUNDAMENTAL FREQUENCY GENERATOR

- The f0 algorithm requires a specification of the number of syllables in each word, the location of stressed syllable within the word and information concerning word boundaries

THE FUNDAMENTAL FREQUENCY GENERATOR

- The f0 algorithm requires a specification of the number of syllables in each word, the location of stressed syllable within the word and information concerning word boundaries
- Output

THE FUNDAMENTAL FREQUENCY GENERATOR

- The f0 algorithm requires a specification of the number of syllables in each word, the location of stressed syllable within the word and information concerning word boundaries
- Output
- The output is target values for midpoint and for peak point at either left or right boundary of stressed vowels in content words

THE O'SHAUGHNESSY FUNDAMENTAL FREQUENCY ALGORITHM

THE O'SHAUGHNESSY FUNDAMENTAL FREQUENCY ALGORITHM

- The algorithm consists of two levels
  - The high level system uses syntactic information to sketch the contour
  - The low level system uses information generated by the high level system and additional phonemic data to detail the contour

The High Level System

THE HIGH LEVEL SYSTEM

- The high level system predicts a superposed f0 contour by taking into consideration sentence type, clause contour, phrase contour, and individual word contour

THE HIGH LEVEL SYSTEM

- The high level system predicts a superposed f0 contour by taking into consideration sentence type, clause contour, phrase contour, and individual word contour
- Sentence Type
  - Two global-level tunes are assigned depending upon sentence type
  - Tune A is used for declaratives and causes a linear falling f0 trend in the clause it is assigned to and a sharp fall in the last content word in the clause and on those words following it
  - Tune B is used for yes/no questions and causes a rise followed by a relatively flat f0 trend and a sharp terminal rise

The High Level System

THE HIGH LEVEL SYSTEM

- Clause Contour
  - A sharp rise is stipulated at the beginning of a syntactic unit and a sharp fall at the end

THE HIGH LEVEL SYSTEM

- Clause Contour
  - A sharp rise is stipulated at the beginning of a syntactic unit and a sharp fall at the end
- Phrase Contour
  - In phrases containing two or more content words, an initial f0 rise is assigned at the first content word and a final f0 fall begins on the last content word
  - If the phrase is non-final, a continuation rise is placed on the last syllable of the last word

The High Level System

THE HIGH LEVEL SYSTEM

- Word Contour
  - Sharp rise and fall on the first and last content words in a phrase
  - A rise-fall contour is described on the stressed syllable of each content word
  - These excursions reflect the desire of the speaker to have listeners understand the less predictable words in a sentence which are also those words carrying which carry the most information
  - The amount of f0 movement depends on its rank in the order of pos of content words and upon the number of syllables in the words

THE HIGH LEVEL SYSTEM

THE HIGH LEVEL SYSTEM

- Prosodic Indicators

THE HIGH LEVEL SYSTEM

- Prosodic Indicators
- A set of prosodic indicators is passed from the high level system to low level system
  - Accent number: 0 for one syllable articles to 11+n for n syllable words
  - Position: indicating the position of a word in a phrase
  - Amount of continuation rise: words ending a non-final phrase are given a value
  - Level: applies to words in noun phrases not containing conjugations. This number signifies either the f0 level is to rise or the level should drop on that word

The High Level System

The High Level System

- Prosodic Indicators

THE HIGH LEVEL SYSTEM

- Prosodic Indicators
- A set of prosodic indicators is passed from the high level system to low level system
  - Accent number: 0 for one syllable articles to 11+n for n syllable words
  - Position: indicating the position of a word in a phrase
  - Amount of continuation rise: words ending a non-final phrase are given a value
  - Level: applies to words in noun phrases not containing conjugations. This number signifies either the f0 level is to rise or the level should drop on that word

THE HIGH LEVEL SYSTEM

THE HIGH LEVEL SYSTEM

- Prosodic Indicators

THE HIGH LEVEL SYSTEM

- Prosodic Indicators
- A set of prosodic indicators is passed from the high level system to low level system
  - Accent number: 0 for one syllable articles to 11+n for n syllable words
  - Position: indicating the position of a word in a phrase
  - Amount of continuation rise: words ending a non-final phrase are given a value
  - Level: applies to words in noun phrases not containing conjugations. This number signifies either the f0 level is to rise or the level should drop on that word
  - Tune: a tune value is defined on each word and is nonzero on the words ending a clause

The Low Level System

THE LOW LEVEL SYSTEM

- This level reflects the effects of phonemics, lexical stress and the number of syllables of the words in the utterance

THE LOW LEVEL SYSTEM

- This level reflects the effects of phonemics, lexical stress and the number of syllables of the words in the utterance
- The number of syllables is used in determining the height of the peak on lexically stressed syllables

THE LOW LEVEL SYSTEM

- This level reflects the effects of phonemics, lexical stress and the number of syllables of the words in the utterance
- The number of syllables is used in determining the height of the peak on lexically stressed syllables
- Higher peaks are also placed on two lexically stressed syllables if they are separated by unstressed syllables, the height of the peaks being dependent upon the number of intervening unstressed syllables

THE LOW LEVEL SYSTEM

THE LOW LEVEL SYSTEM

- The f0 pattern is also affected by phonemics e.g. unvoiced consonants at the beginning of a stressed syllable also cause the contour to fall rather than rise into the contour of the stressed vowel

The Low Level System

THE LOW LEVEL SYSTEM

- The algorithm first sets peaks on the lexically stressed syllables

THE LOW LEVEL SYSTEM

- The algorithm first sets peaks on the lexically stressed syllables
- Falls and rises are then assgined around peaks.

THE LOW LEVEL SYSTEM

- The algorithm first sets peaks on the lexically stressed syllables
- Falls and rises are then assgined around peaks.
- Continuation rises are added to the last syllable of most non-final phrases

THE LOW LEVEL SYSTEM

- The algorithm first sets peaks on the lexically stressed syllables
- Falls and rises are then assgined around peaks.
- Continuation rises are added to the last syllable of most non-final phrases
- Sentence-final words are given rises and falls depending upon their tune

THE LOW LEVEL SYSTEM

- The algorithm first sets peaks on the lexically stressed syllables
- Falls and rises are then assgined around peaks.
- Continuation rises are added to the last syllable of most non-final phrases
- Sentence-final words are given rises and falls depending upon their tune
- Finally the f0 contour is completed by specifying the amount of fall on other unstressed syllables

THE PHONETIC COMPONENT

THE PHONETIC COMPONENT

- Input
  - The PHONET module accepts an array of phonetic names and a segmental stress feature, segmental duration and two fundamental frequency targets for each phone

THE PHONETIC COMPONENT

- Input
  - The PHONET module accepts an array of phonetic names and a segmental stress feature, segmental duration and two fundamental frequency targets for each phone
- Output
  - The output is values for 20 synthesizer control parameters every 5 msec

THE PHONETIC COMPONENT

THE PHONETIC COMPONENT

- Stored Prosodic Synthesis

THE PHONETIC COMPONENT

- Stored Prosodic Synthesis
- The phonetic component and synthesizer can be used stand alone mode in which the phonetic segment string, durations, and fundamental frequency contour specification are manually given as input

THE PHONETIC COMPONENT

- Stored Prosodic Synthesis
- The phonetic component and synthesizer can be used stand alone mode in which the phonetic segment string, durations, and fundamental frequency contour specification are manually given as input
- Highly intelligible speech is generated is these parameters are obtained by analyzing natural utterance

THE PHONETIC COMPONENT

THE PHONETIC COMPONENT

- Structure of PHONET

THE PHONETIC COMPONENT

- Structure of PHONET
- It includes a large array of target values for various control parameters for each of about 60 phonetic segment types

THE PHONETIC COMPONENT

- Structure of PHONET
- It includes a large array of target values for various control parameters for each of about 60 phonetic segment types
- Smoothing between target values depends on time constants that are computed by rule as well as depending on the parameter value assigned to the time of the segment boundary

THE PHONETIC COMPONENT

- Structure of PHONET
- It includes a large array of target values for various control parameters for each of about 60 phonetic segment types
- Smoothing between target values depends on time constants that are computed by rule as well as depending on the parameter value assigned to the time of the segment boundary
- These constants are determined by rules that involve features of current, previous and next phonetic segment

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- A database was recorded and analyzed in order to develop e new consonant-vowel synthesis rules

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- A database was recorded and analyzed in order to develop e new consonant-vowel synthesis rules
- The speech was lowpass filtered and sampled

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- A database was recorded and analyzed in order to develop e new consonant-vowel synthesis rules
- The speech was lowpass filtered and sampled
- LP spectra were computed at a number of hand-selected locations in a syllable

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- A database was recorded and analyzed in order to develop e new consonant-vowel synthesis rules
- The speech was lowpass filtered and sampled
- LP spectra were computed at a number of hand-selected locations in a syllable
- The waveform was pre-emphasized and windowed followed by 11-pole LP analysis

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- Spectral samples were obtained
  - During the consonantal steady state (or at burst onset for plosive)
  - At voicing onset (or early in the CV transition for voiced consonants)
  - Shortly after the end of CV transition

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- Spectral samples were obtained
  - During the consonantal steady state (or at burst onset for plosive)
  - At voicing onset (or early in the CV transition for voiced consonants)
  - Shortly after the end of CV transition
- Formant frequencies were estimated by locating the peaks in a LP spectrum

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- Spectral samples were obtained
  - During the consonantal steady state (or at burst onset for plosive)
  - At voicing onset (or early in the CV transition for voiced consonants)
  - Shortly after the end of CV transition
- Formant frequencies were estimated by locating the peaks in a LP spectrum
- Formant motions were plotted every 10msec during voiced portions of syllables

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

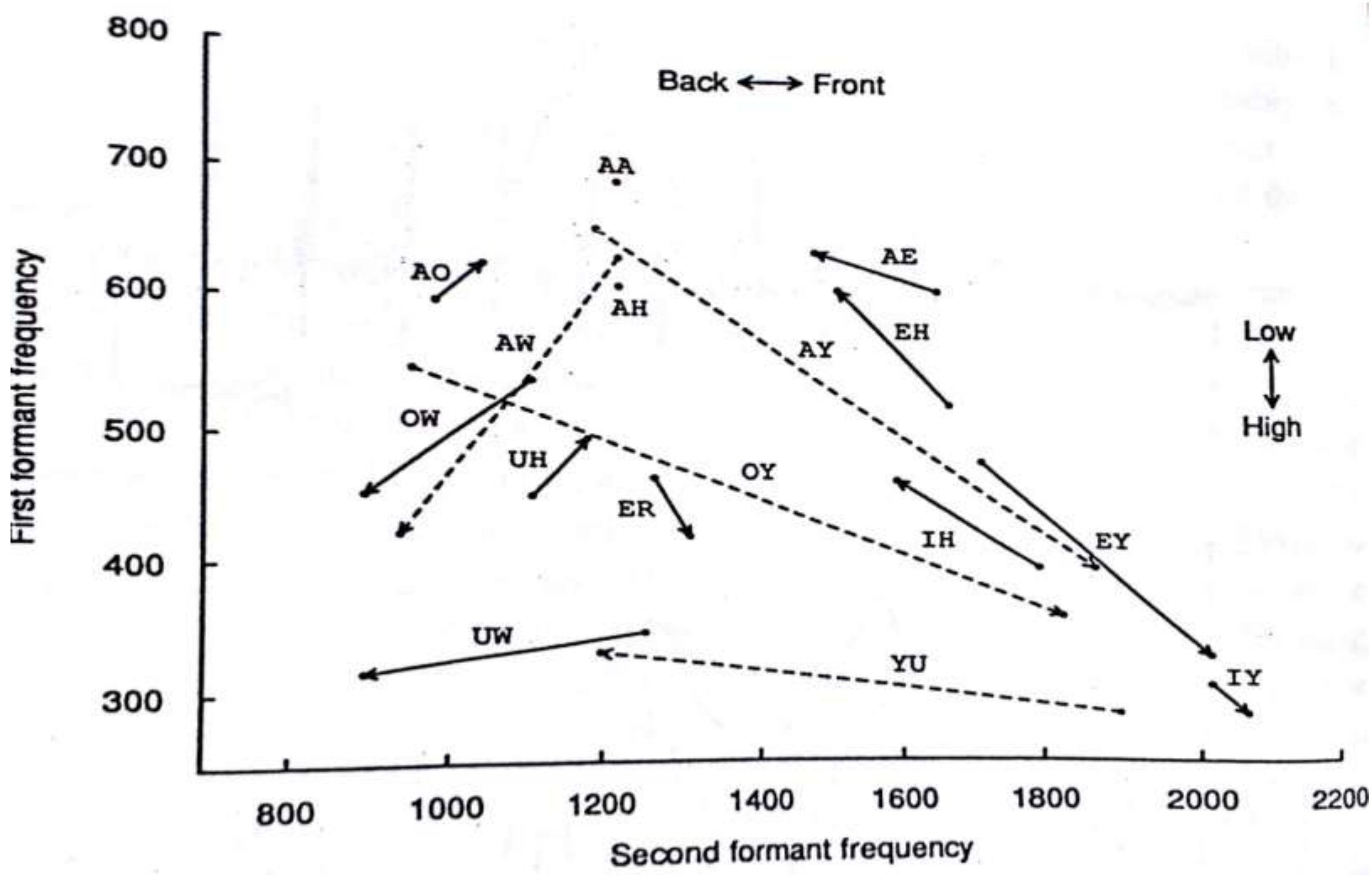
- Spectral samples were obtained
  - During the consonantal steady state (or at burst onset for plosive)
  - At voicing onset (or early in the CV transition for voiced consonants)
  - Shortly after the end of CV transition
- Formant frequencies were estimated by locating the peaks in a LP spectrum
- Formant motions were plotted every 10msec during voiced portions of syllables
- Intensity and f0 were also estimated and plotted as a function of time

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- Figure presents first and second formant frequency trajectories of sixteen vowel nuclei as averaged across all consonantal environments for the designated talker

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

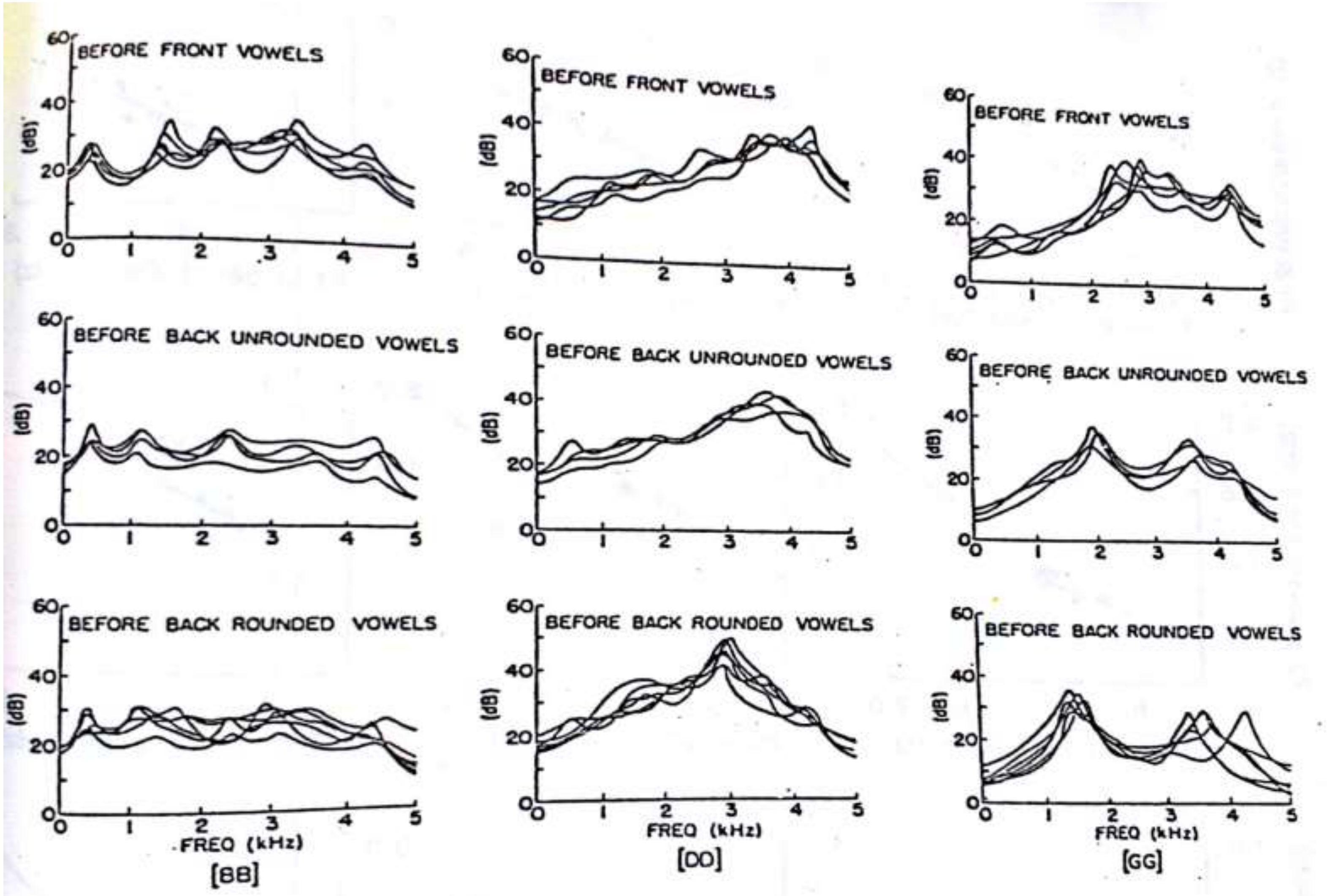


SYNTHESIS BY ANALYSIS OF CV SYLLABLES

SYNTHESIS BY ANALYSIS OF CV SYLLABLES

- Analysis of consonants revealed two major conclusions
  - Some consonants, particularly obstruents, take on significantly different characteristics depending on whether the following vowel is a front vowel, a back unrounded vowel or a back rounded vowel
  - Within each set of vowels, the spectra associated with each consonant are surprisingly invariant and formant transitions into the vowel obey a modified locus equation

SYNTHESIS BY ANALYSIS OF CV SYLLABLES



SYNTHESIS STRATEGY

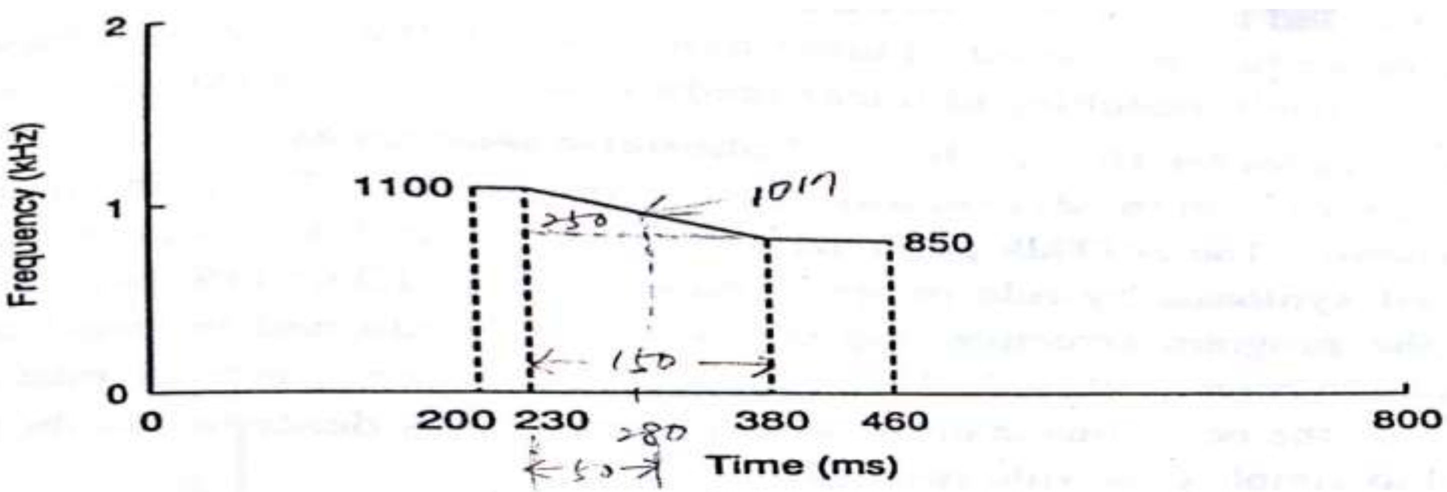
SYNTHESIS STRATEGY

- For formant frequency motions in CV syllables
  - The vowel is first defined in terms of straight line segment
  - The formant values associated with the consonant and the consonant-vowel transition are imposed using straight line segments
  - A locus theory equation is used to determine formant values at the CV boundary

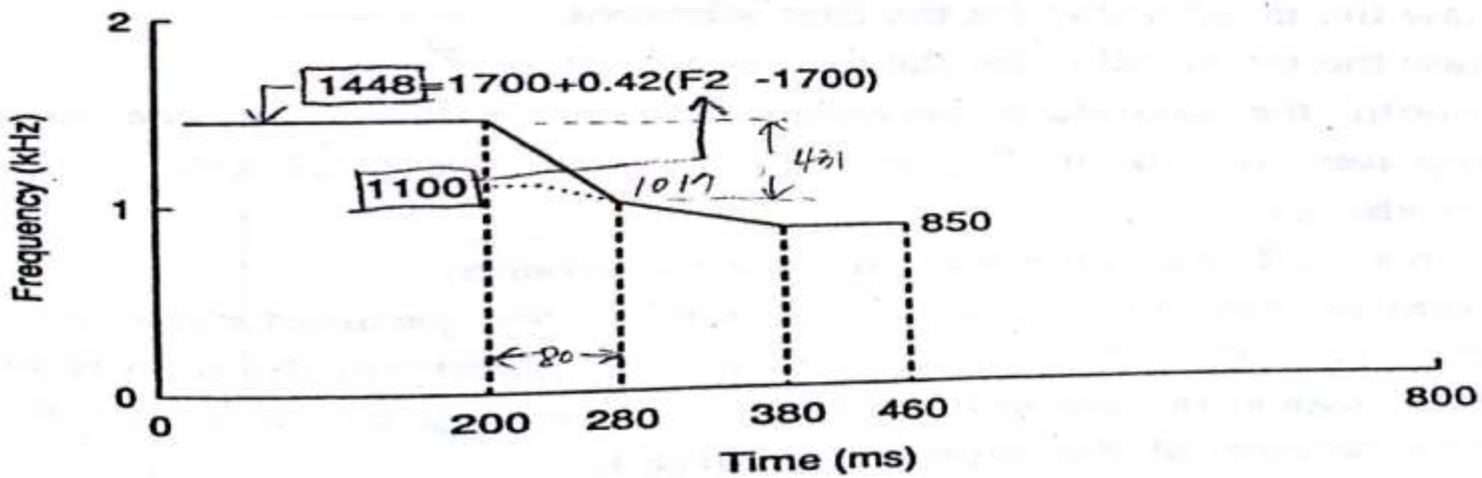
SYNTHESIS STRATEGY

- For formant frequency motions in CV syllables
  - The vowel is first defined in terms of straight line segment
  - The formant values associated with the consonant and the consonant-vowel transition are imposed using straight line segments
  - A locus theory equation is used to determine formant values at the CV boundary
- Many of the remaining synthesis parameters such as formant bandwidths and formant amplitudes in frication spectra must be determined by trial and error comparisons of synthetic and natural LP spectra

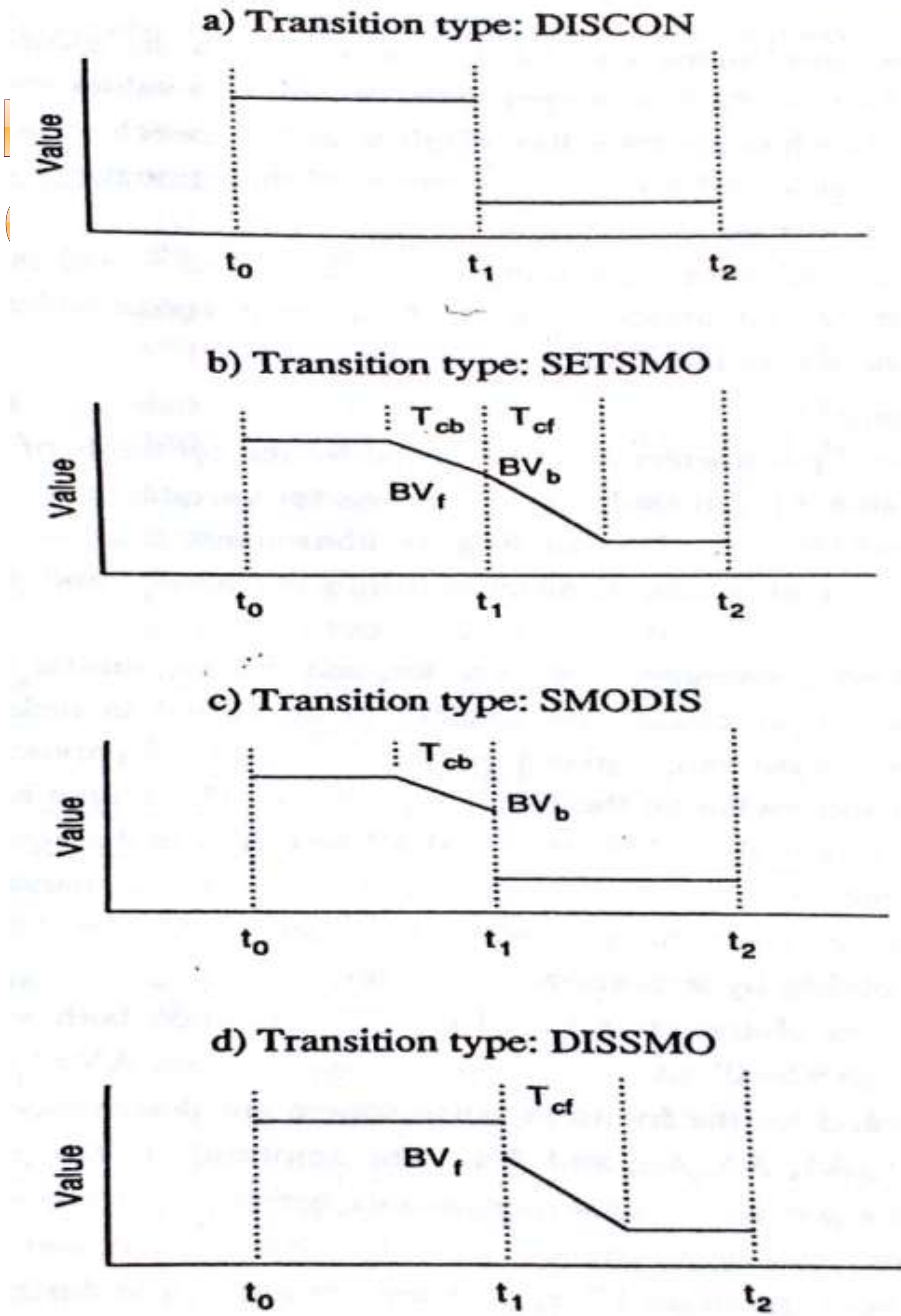
SYNTHESIS STRATEGY



Second formant transition for GG OW

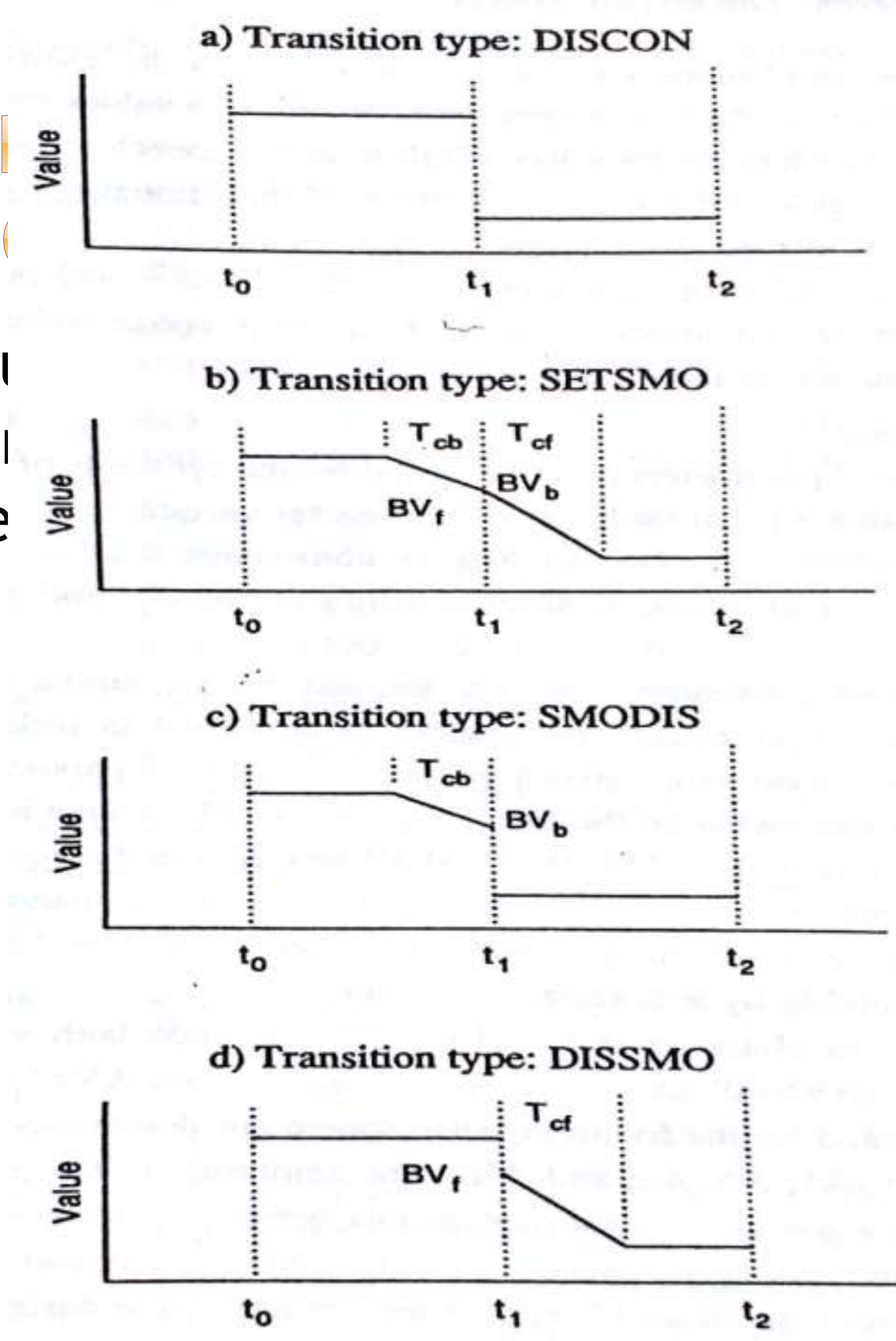


GENE THESIS
S

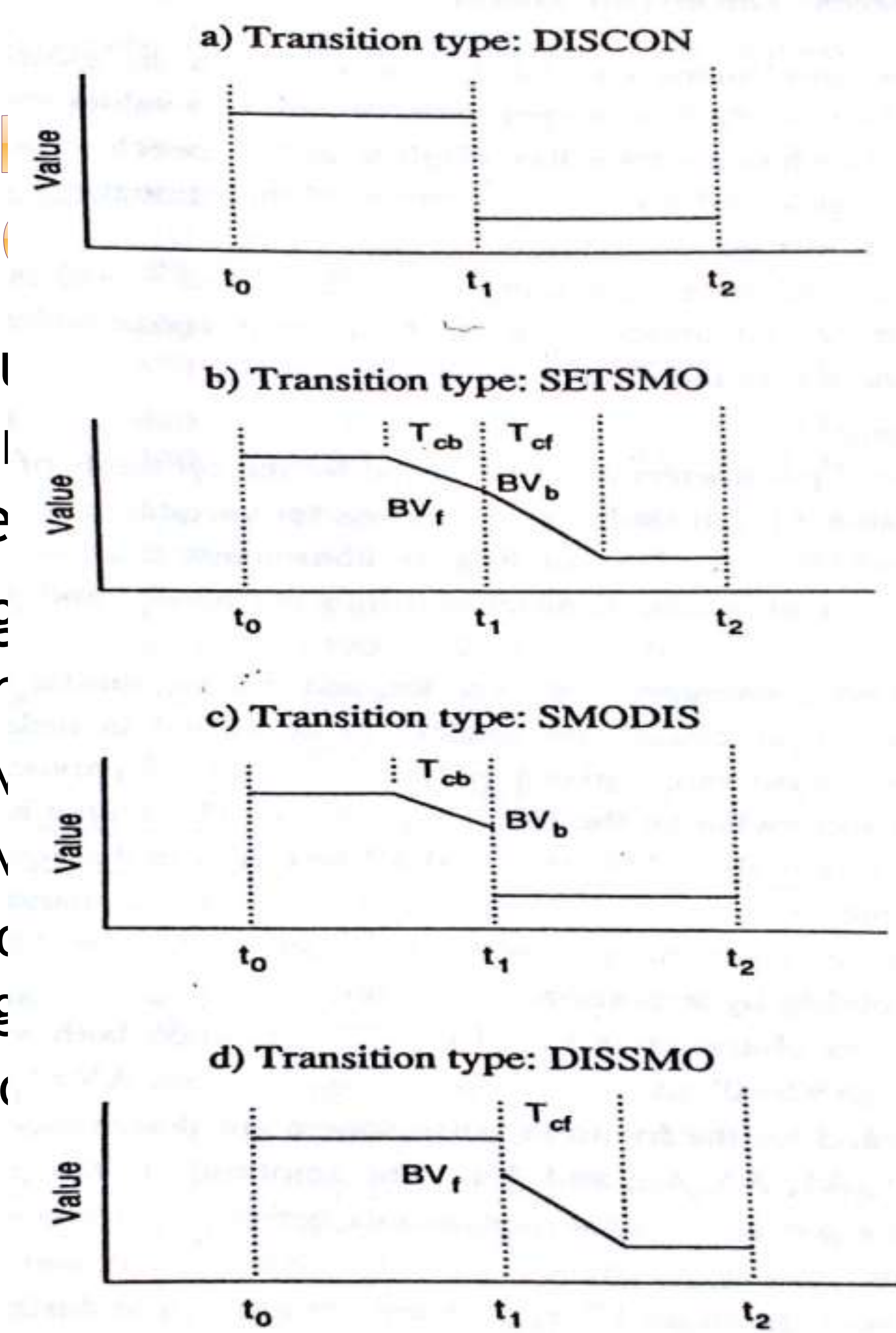


GENE… THESIS
(…S

- The r… fers from the li… described above



GENE[…] THESIS
([…]S



- The ru[…]fers from the li[…]described above[…]
- The g[…]g control param[…]
  - Draw[…]
  - Draw[…]
  - Smo[…] segments using[…]
  - Go t[…]egments

The Klatt Formant Synthesizer (in brief)

THE KLATT FORMANT SYNTHESIZER (IN BRIEF)

- A brief review of the last two components of MITalk (i.e CWTRAN and COEWAV) which simulate a formant synthesizer is given

Cascade vs. Parallel

CASCADE VS. PARALLEL

- Parallel Formants

CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum

CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum
  - Certain vowels are difficult to synthesize using parallel formants

CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum
  - Certain vowels are difficult to synthesize using parallel formants
- Cascaded Formants

CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum
  - Certain vowels are difficult to synthesize using parallel formants
- Cascaded Formants
  - No individual formant amplitude control is required

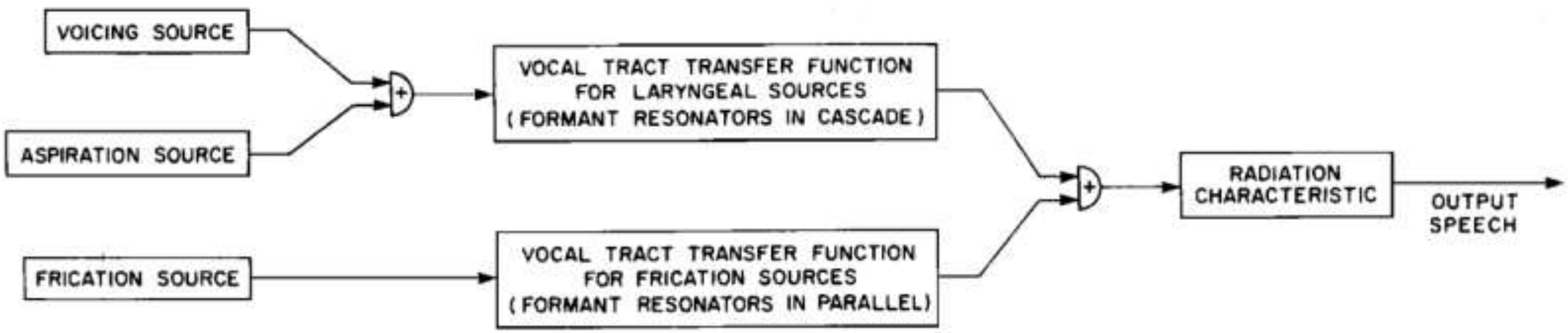
CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum
  - Certain vowels are difficult to synthesize using parallel formants
- Cascaded Formants
  - No individual formant amplitude control is required
  - Cascade is more accurate model for semivowels

CASCADE VS. PARALLEL

- Parallel Formants
  - Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak in the output spectrum
  - Certain vowels are difficult to synthesize using parallel formants
- Cascaded Formants
  - No individual formant amplitude control is required
  - Cascade is more accurate model for semivowels
  - For fricatives and plosives bursts still parallel formants are required

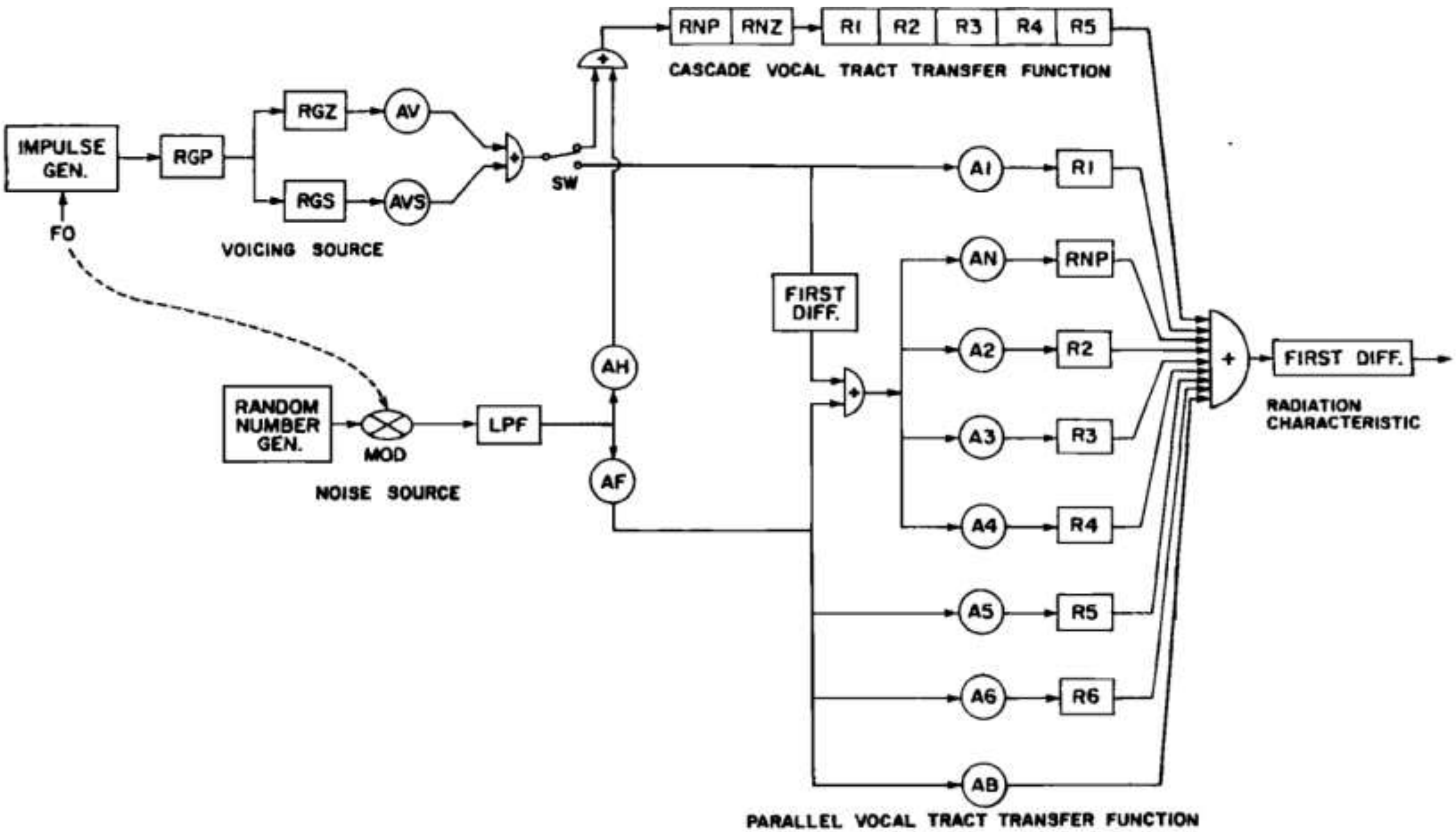
CASCADE VS. PARALLEL



(A) CASCADE / PARALLEL FORMANT CONFIGURATION

(B) SPECIAL-PURPOSE ALL-PARALLEL FORMANT CONFIGURATION

SYNTHESIZER BLOCK DIAGRAM



DIGITAL RESONATORS

DIGITAL RESONATORS

- The basic building block of the synthesizer is a digital resonator

$$y(\mathrm{n}) = \mathrm{A}.\mathrm{x}(\mathrm{n}) + \mathrm{B}.\mathrm{y}(\mathrm{n}-1) + C.\mathrm{y}(\mathrm{n}-2)$$

DIGITAL RESONATORS

- The basic building block of the synthesizer is a digital resonator

$$y(\mathrm{n}) = \mathrm{A.x(n)} + \mathrm{B.y(n-1)} + C.\mathrm{y(n-2)}$$

- The constants A, B, and C are related to the resonant frequency F and the bandwidth BW of a resonator by the impulse-invariant transformation as

$$A = 1 - B - C$$

$$B = 2e^{-\frac{2\pi B}{f_s}} \cos(\frac{2\pi f}{f_s})$$

$$C = -e^{-\frac{2\pi B}{f_s}}$$

DIGITA

- The ba … is
  a digita
- The co … he
  resonar … BW
  of a … nt
  transfo

$$H(z) = \frac{1}{(1 - z_1 z^{-1})(1 - z_1^* z^{-1})}$$

$$z_1 = \gamma e^{j\omega_0}$$

$$H(z) = \frac{1}{(1 - \gamma e^{j\omega_0} z^{-1})(1 - \gamma e^{-j\omega_0} z^{-1})}$$

$$H(z) = \frac{1}{1 - 2\gamma \cos\omega_0 z^{-1} + \gamma^2 z^{-2}}$$

$$H(z) = \frac{1}{1 - b z^{-1} - c z^{-2}}$$

$$b = 2\gamma \cos\omega_0 = 2 e^{-\pi BWT} \cos(2\pi f_0 T)$$

$$c = -\gamma^2 = - e^{-2\pi BWT}$$

For unity magnitude at f_0

$$a = 1 - b - c$$

```
T=0.0001;
f0=1000;
BW=200;
b=2*exp(-pi*BW*T)*cos(2*pi*f0*T);
c=-exp(-2*pi*BW*T);
a=1-b-c;
f=0:5000;
w=2*pi*(f)*T;

freqresp0=abs(a./(1-b.*exp(-1j.*w)-c.*exp(-1j*2.*w)));
plot(f,freqresp0) % magnetude greater than unity at f0

freqz([a], [1, -b, -c]) % magnetude greater than unity

w=2*pi*(f0- BW/2)*T;
freqresp1=abs(a/(1-b*exp(-1j*w)-c*exp(-1j*2*w)));
w=2*pi*(f0)*T;
freqresp0=abs(a/(1-b*exp(-1j*w)-c*exp(-1j*2*w)));
freqresp1/freqresp0 %#ok<NOPTS> %answer is 0.7395
```

SOUND SOURCES

SOUND SOURCES

- Six kind of sources
  - Normal Voicing
  - Quasi-Sinusoidal Voicing
  - Normal Frication
  - Amplitude Modulated Frication
  - Normal Aspiration
  - Amplitude Modulated Aspiration

FORMANT FREQUENCIES

FORMANT FREQUENCIES

- Formant frequency values are determined by the detailed shape of the vocal tract

FORMANT FREQUENCIES

- Formant frequency values are determined by the detailed shape of the vocal tract
- The frequencies of the lowest three formants vary substantially with changes to articulation

FORMANT FREQUENCIES

- Formant frequency values are determined by the detailed shape of the vocal tract
- The frequencies of the lowest three formants vary substantially with changes to articulation
- Higher frequency resonators help to shape the overall spectrum, but otherwise contribute little to intelligibility for vowels

FORMANT FREQUENCIES

- Formant frequency values are determined by the detailed shape of the vocal tract
- The frequencies of the lowest three formants vary substantially with changes to articulation
- Higher frequency resonators help to shape the overall spectrum, but otherwise contribute little to intelligibility for vowels
- The particular values chosen for the fourth and fifth formant produce an energy concentration around 3 to 3.5 kHz and a rapid falloff in spectral energy above about 4 kHz, which is a pattern typical of many talkers

Formant Bandwidths

FORMANT BANDWIDTHS

- Results indicate that bandwidths vary by a factor of 2 or more as a function of the particular phonetic segment being spoken

FORMANT BANDWIDTHS

- Results indicate that bandwidths vary by a factor of 2 or more as a function of the particular phonetic segment being spoken
- The primary perceptual effect of a bandwidth change is an increase or decrease in the effective intensity of a formant energy concentration

FORMANT BANDWIDTHS

- Results indicate that bandwidths vary by a factor of 2 or more as a function of the particular phonetic segment being spoken
- The primary perceptual effect of a bandwidth change is an increase or decrease in the effective intensity of a formant energy concentration
- Bandwidth variation is small enough that all formant bandwidths might be held constant in some applications

Radiation Characteristics

RADIATION CHARACTERISTICS

- The sound pressure measured directly in front of and about a meter from the lips is proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r, the distance from the lips

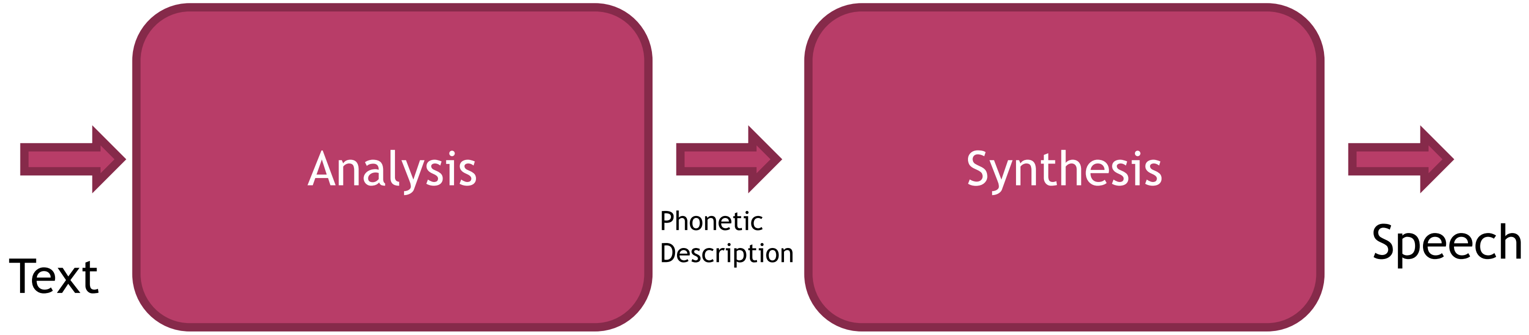
RADIATION CHARACTERISTICS

- The sound pressure measured directly in front of and about a meter from the lips is proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r, the distance from the lips
- The transformation is simulated in the synthesizer by taking the first difference of lip-nose volume velocity

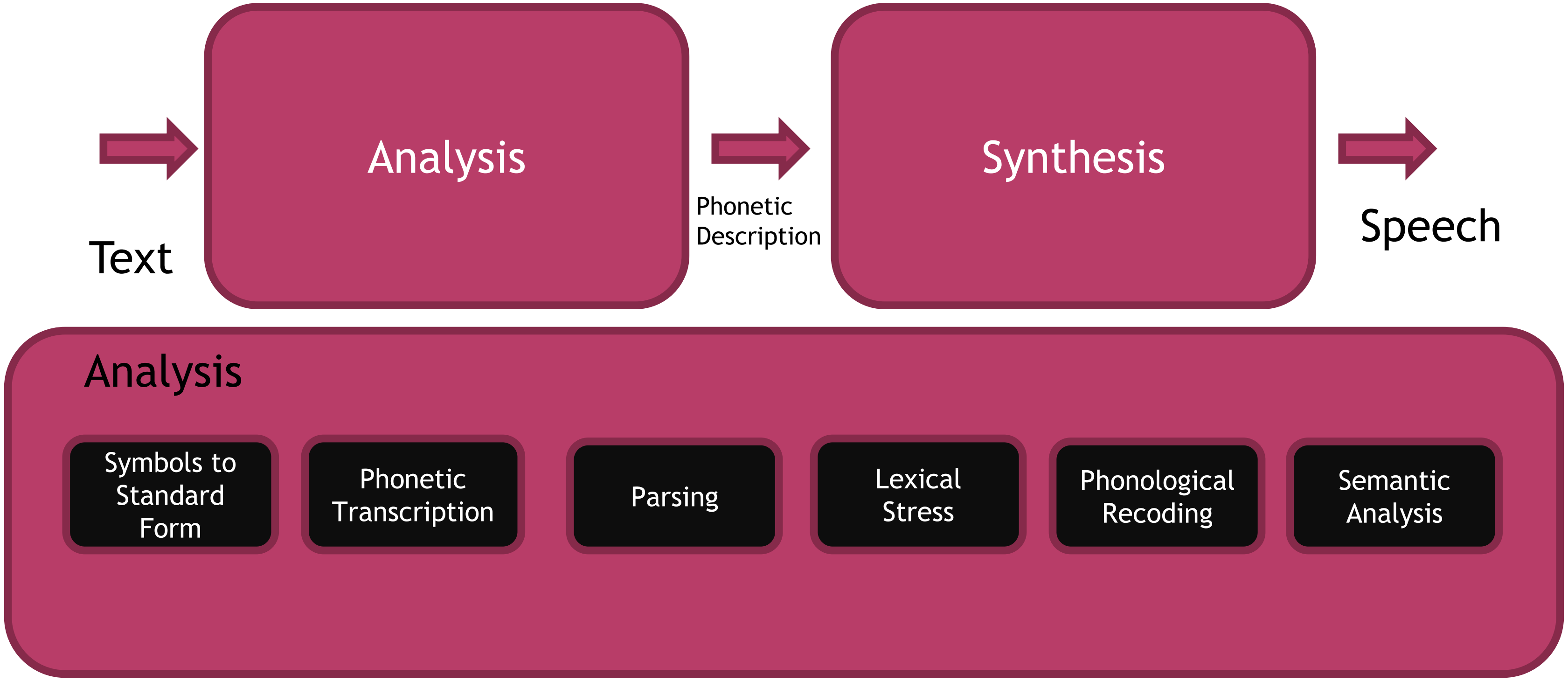
$$p(n) = u(n) - u(n-1)$$

FUNCTIONAL OUTLINE OF MITALK

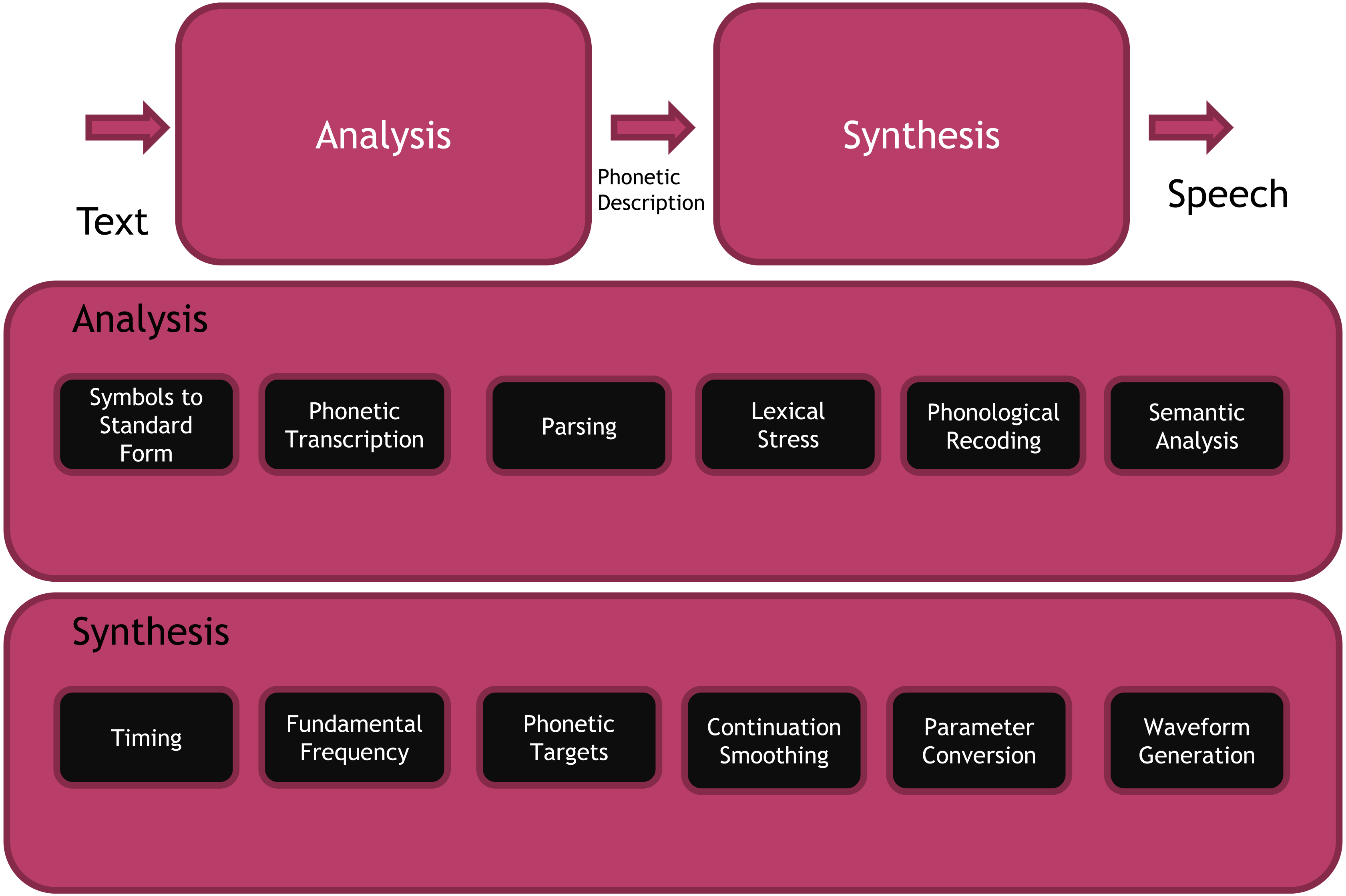
FUNCTIONAL OUTLINE OF MITALK



FUNCTIONAL OUTLINE OF MITALK



FUNCTIONAL OUTLINE OF MITALK



Thank You