Disclaimer

- The material provided in this document is not my original work and is a summary of some one else's work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
- Downloaded from najeebkhan.github.io

Software for Cascade Parallel Formant Synthesizer

Presented By
Najeeb
2012 – 5 – 10

Outline

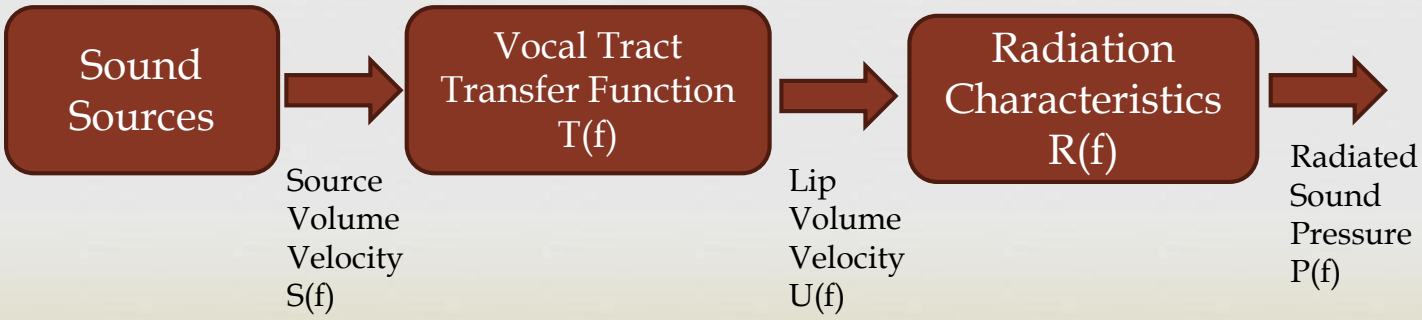
- Introduction
- Sources of Sound
- Vocal Tract Transfer Functions
- Radiation Characteristics
- Synthesis Strategy
- Conclusions

Introduction

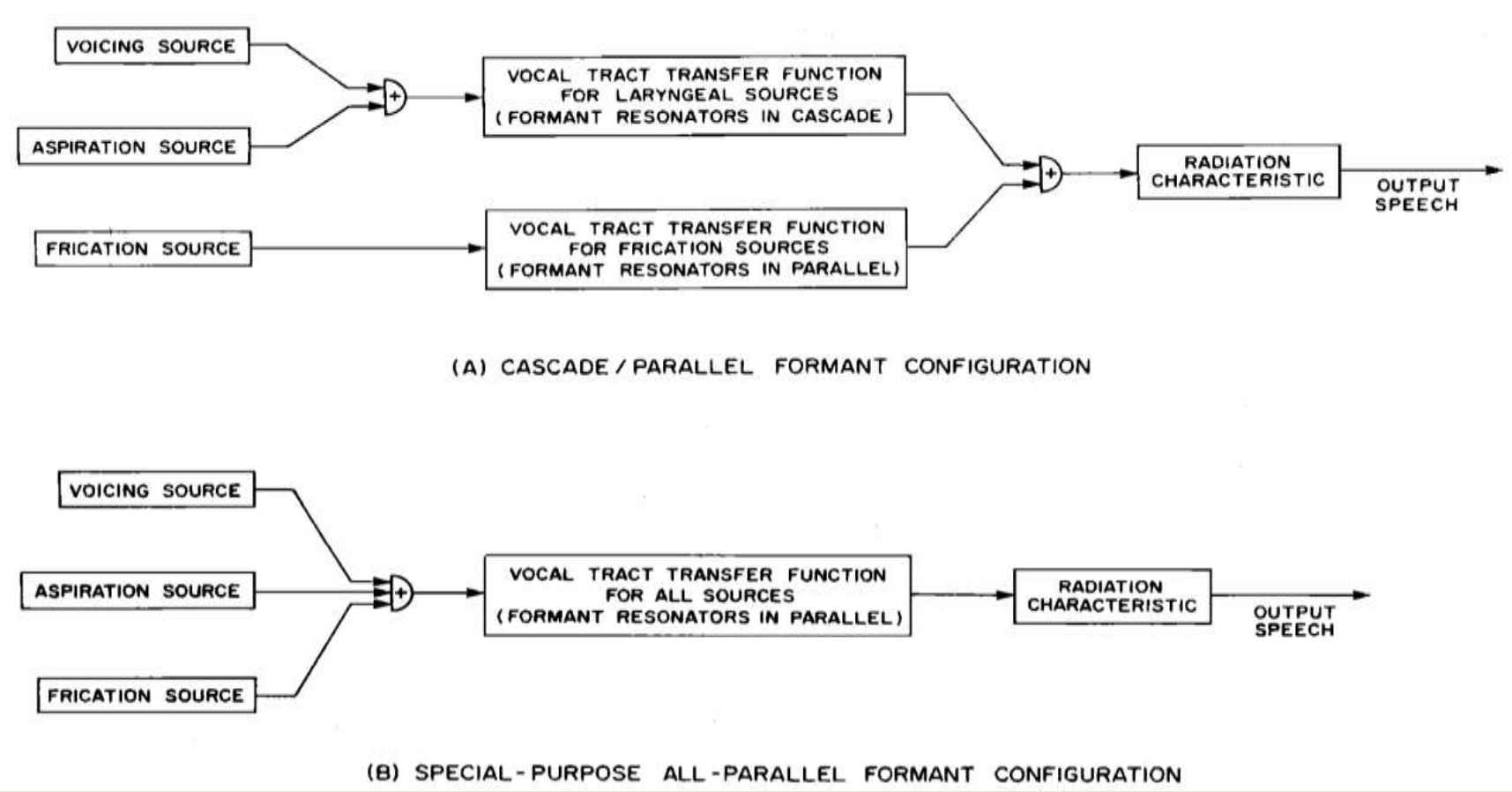
- A software formant synthesizer is described that can generate synthetic speech using a Computer.
- The synthesizer design is based on an acoustic theory of speech production

Acoustic Theory of Speech Production

- One or more sound sources are activated by the build up of lung pressure
- Each sound source excites the vocal tract which acts as a resonating system analogous to an organ pipe
- The spectrum of the sound pressure P(f) that would be recorded some distance from the lips of the talker, is related to U(f) by R(f)



Cascade Vs. Parallel



Sampling Rate

- Most of the sound energy of speech is contained in frequencies between about 80 and 8000Hz
- Intelligibility is not measurably changed if the energy in frequencies above about 5000Hz is removed
- The default sampling rate for the synthesizer is 10KHz
- Control parameters are updated every 5ms, i.e. every 50 samples

Digital Resonators

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

$$y(n) = A.x(n) + B.y(n-1) + C.y(n-2)$$

- 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

$$A = 1 - B - C$$

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

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

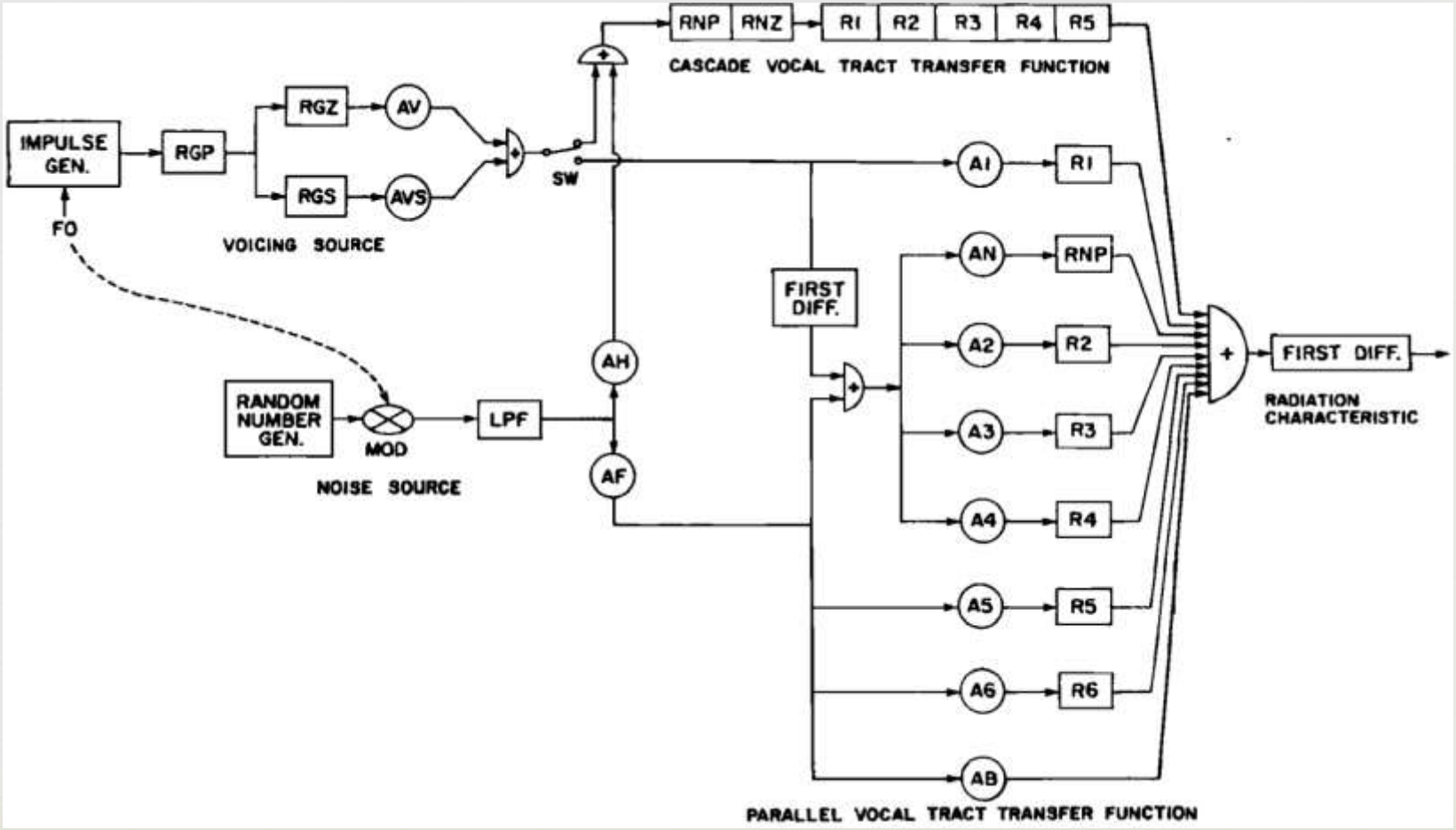
Digital Antiresonator

- Used in synthesizer to
  - Shape the spectrum of the voicing source
  - Simulate the effects of nasalization in the cascade model of the vocal tract transfer function
- The Antiresonator is given by

$$y(\mathrm{n}) = a.\mathrm{x}(\mathrm{n}) + b.\mathrm{x}(\mathrm{n}-1) + c.\mathrm{x}(\mathrm{n}-2)$$

$$a = 1/A \quad b = -B/\mathrm{A} \quad c = -CA$$

Synthesizer Block Diagram



Sound Sources

- Vibrations of the vocal folds called voicing
- Generation of noise by the rapid flow of air past a narrow constriction
  - The resulting noise is called aspiration if the constriction is located at the level of the vocal folds [h]
  - The resulting noise is called frication if the constriction is located above the larynx [s]
- If voice and noise co-exist
  - Noise is amplitude modulated periodically by the vibration of the vocal folds
- Vocal folds may vibrate without meeting in the midline resulting in nearly sinusoidal waveform

Sound Sources

- So we have six kind of sources
  - Normal Voicing
  - Quasi-Sinusoidal Voicing
  - Normal Frication
  - Amplitude Modulated Frication
  - Normal Aspiration
  - Amplitude Modulated Aspiration

Voicing Sources

- F0, AV and AVS are the voicing control parameters
- Impulse train corresponding to normal voicing is generated with amplitude AV
- The number of samples between impulses T0 is determined by Sampling rate/F0

Normal Voicing

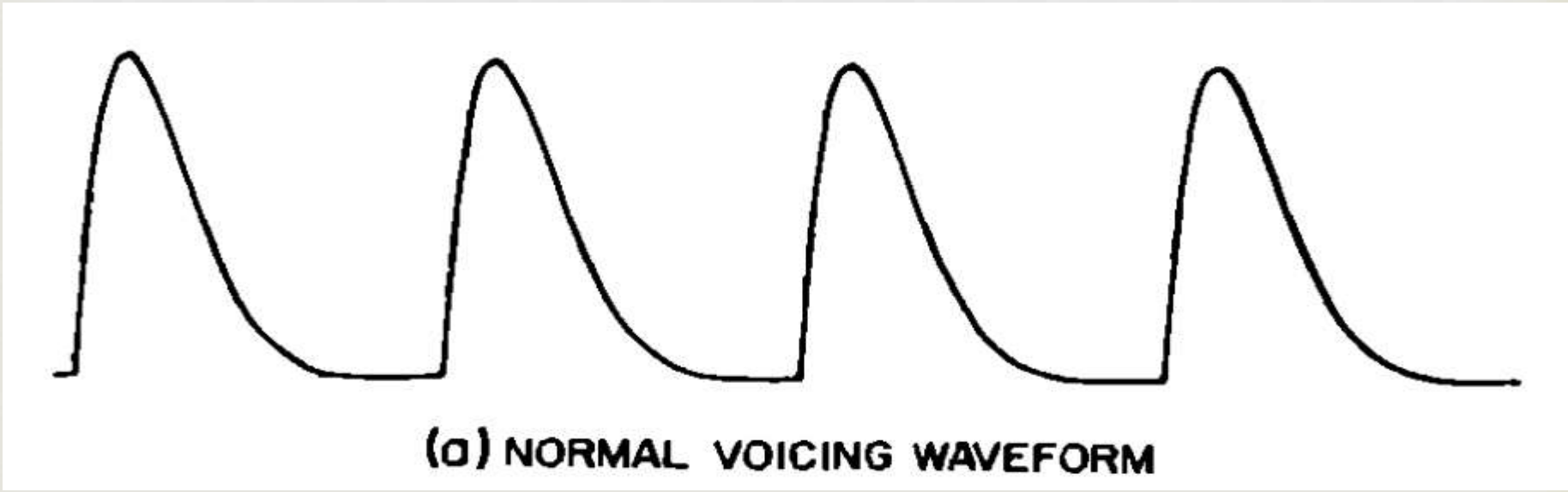
- The train of impulses is sent through a lowpass filter, RGP, to produce a smooth waveform that resembles a typical glottal volume velocity waveform

$$FGP = 0 \quad BGP = 100Hz$$

- -12dB/octave above 50Hz
- The glottal Antiresonator RGZ is used to modify the detailed shape of the spectrum of the voicing source for particular individuals.

Contd…

- The waveform produced does not have the same phase spectrum as a typical glottal pulse nor does it contain spectral zeros of the kind that appear in natural voicing



Quasi Sinusoidal Source

- AVS determines the amount of smoothed voicing generated during voiced fricatives, voiced aspirates and the voice bars present in intervocalic voiced plosives
- RGS is used to filter the glottal pulses

$$FGS = 0 \quad BGS = 200Hz$$

(b) SMOOTHED VOICING WAVEFORM

Frication Source

- Noise source is simulated in the synthesizer by a pseudo-random number generator, a modulator, amplitude AF and a -6dB/octave LPF.
- The spectrum of the noise source should be approximately flat
- The amplitude distribution should be Gaussian
- The output of the random number generator is amplitude modulated whenever the AV and F0 are non-zero.

Aspiration Source

- Aspiration noise is essentially the same as frication noise, except that it is generated in the larynx.
- In the cascade synthesizer configuration, aspiration noise is sent through the cascade vocal tract model while fricatives require a parallel vocal tract configuration
- Therefore separate amplitude controls are needed for frication and aspiration in a cascade/parallel configuration

Control of Source Amplitudes

- Parameter values specifying source amplitudes AV, AVS, AF, and AH are adjusted by the user to new values every 5ms
- AV and AVS only have an effect on the synthetic waveform when a glottal impulse is issued
- The noise amplitudes AF and AH are used to interpolate the intensity of the noise sources linearly over the 5-ms (50-sample) interval

Contd…

- A plosive burst involves a more rapid source onset than can be achieved by 5ms linear interpolation
- Therefore, if AF increases by more than 50 dB from its value specified in the previous 5ms segment, AF is changed instantaneously to its new target value

Control Of F0…

- A glottal pulse is issued in the synthesizer at a time specified by one over the value of the fundamental frequency control parameter value extant when the last glottal pulse was issued
- If either AV or F0 is set to zero, no glottal pulse is issued during this 5-ms time interval
- Since the update interval in the synthesizer is set to 5ms, voice onset time can be specified exactly in 5ms steps

Control of Noise Samples in Stimulus Continuum

- A particular brief noise sequence may have greater or lesser total intensity, or a peculiar spectral peak or valley not shared by other samples of noise that are used to generate a set of stimuli varying in voice onset time or burst frequency
- These random fluctuations in noise characteristics can cause some stimuli in a supposed continuum to stand out as different
- Use the same random number sequence in the generation of each member of the continuum by reinitializing the random number function

Vocal Tract Transfer Function

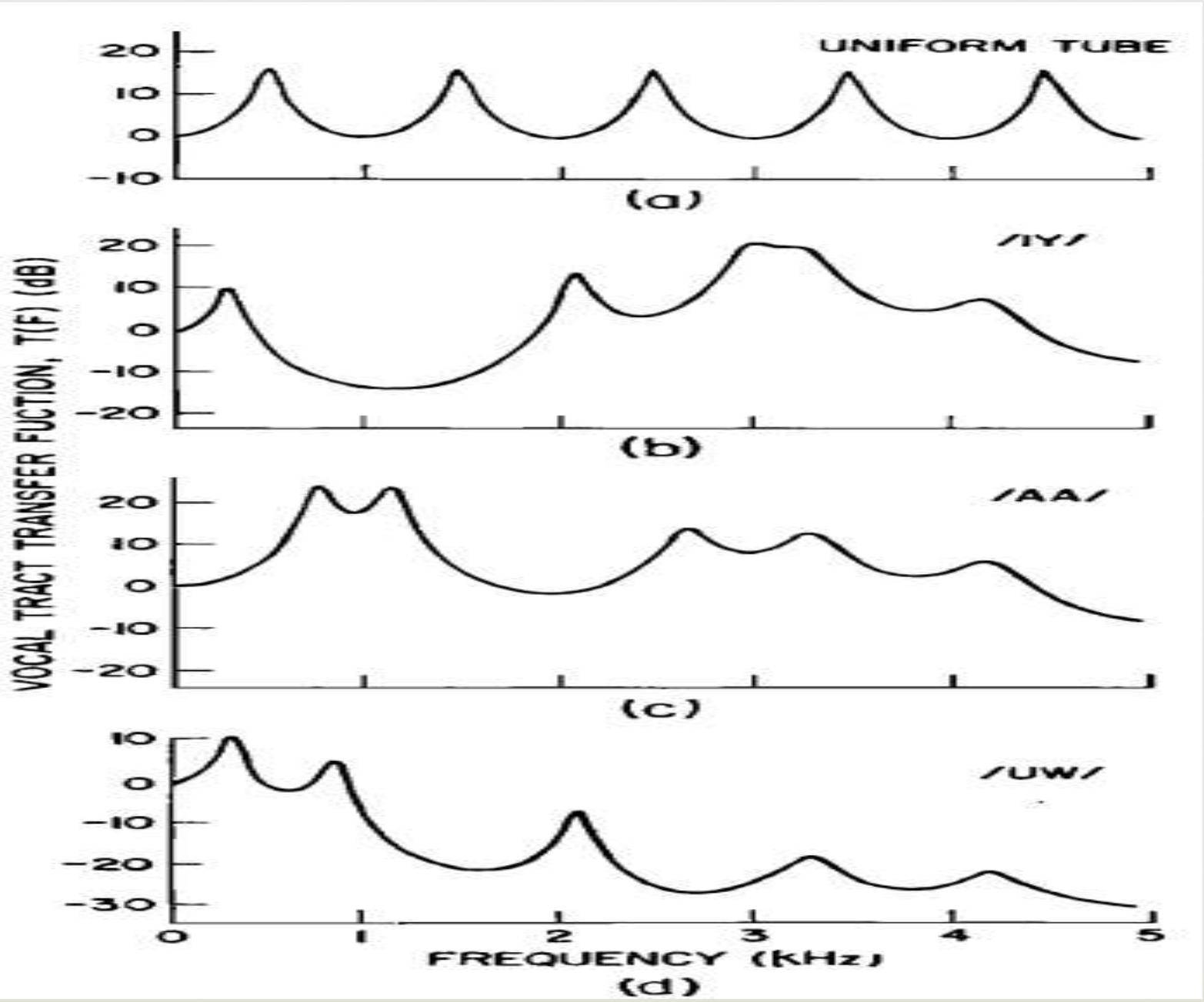
- The acoustic characteristics of the vocal tract are determined by its cross sectional area as a function of distance from the larynx to the lips
- The vocal tract forms a non uniform transmission line whose behavior can be determined for frequencies below about 5 kHz by solving a one dimensional wave equation
- Solutions to the wave equation result in a transfer function that relates samples of the glottal source volume velocity to output volume velocity at the lips

Contd…

- Two different realizations of vocal tract transfer function
  - A cascade configuration of digital resonators, models the resonant properties of the vocal tract whenever the source of sound is within the larynx
  - A parallel configuration of digital resonators and amplitude controls, models the resonant properties of the vocal tract during the production of frication noise

Cascade Vocal Tract Model

- The vocal tract transfer function can be represented in the frequency domain by a product of poles and zeros
- Transfer function contains only about five complex pole pairs and no zeros in the frequency range of interest, as long as the articulation is non-nasalized and the sound source is at the larynx
- The average spacing between formants is equal to the velocity of sound divided by half the wavelength, which works out to be 1000 Hz


UNIFORM TUBE
(a)
/IY/
(b)
/AA/
(c)
/UW/
(d)
VOCAL TRACT TRANSFER FUCTION, T(F) (dB)
FREQUENCY (kHz)
20
10
0
-10
-20
-30
0
1
2
3
4
5

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

- 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

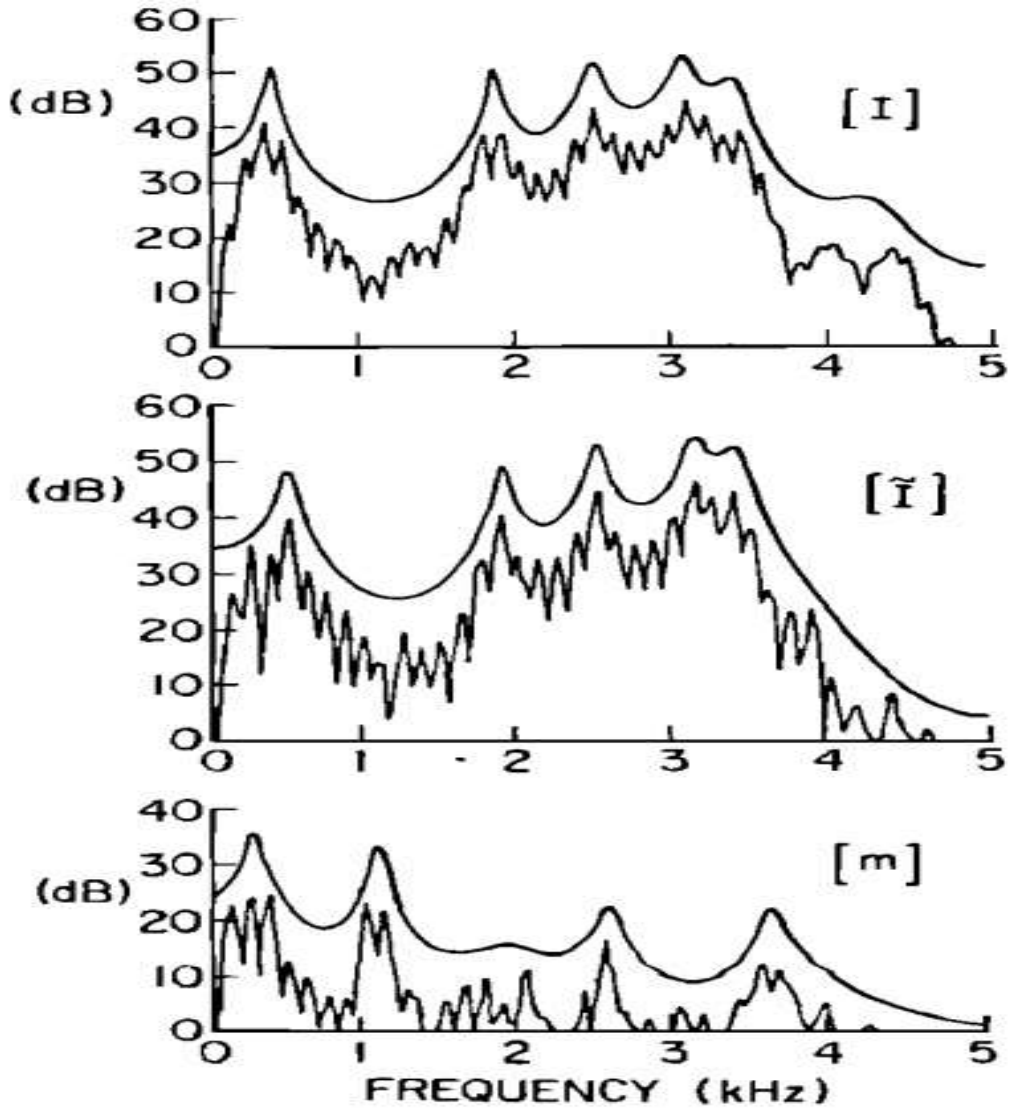


FIG. 10. Spectra are compared of a the vowel [I], the same vowel when nasalized, and a nasal murmur [m], all obtained from the recorded syllable "dim". The nasal murmur and the nasalized [Ĩ] have an extra transfer function pole pair and zero pair near *F*1. The extra peak and valley are not appearant in the linear prediction spectrum, but can be discerned in the pattern of harmonic amplitudes near *F*1 in the discrete Fourier transform spectrum.

Contd…

- Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator RNP and Antiresonator RNZ into the cascade vocal tract model
- The nasal pole frequency FNP can be set to a fixed value of about 270 Hz for all time
- The nasal zero frequency FNZ should also be set to a value of about 270 Hz during non-nasalized sounds, but the frequency of the nasal zero must be increased during the production of nasals and nasalization

Parallel Vocal Tract Model for Frication Sources

- During frication excitation, the vocal tract transfer function contains both poles and zeros
- The effect of transfer-function zeros is twofold
  - They introduce notches in the spectrum
  - They modify the amplitudes of the formants
- The perceptual importance of spectral notches is not great because masking effects of adjacent energy in format peaks limit the delectability of a spectral notch

Contd…

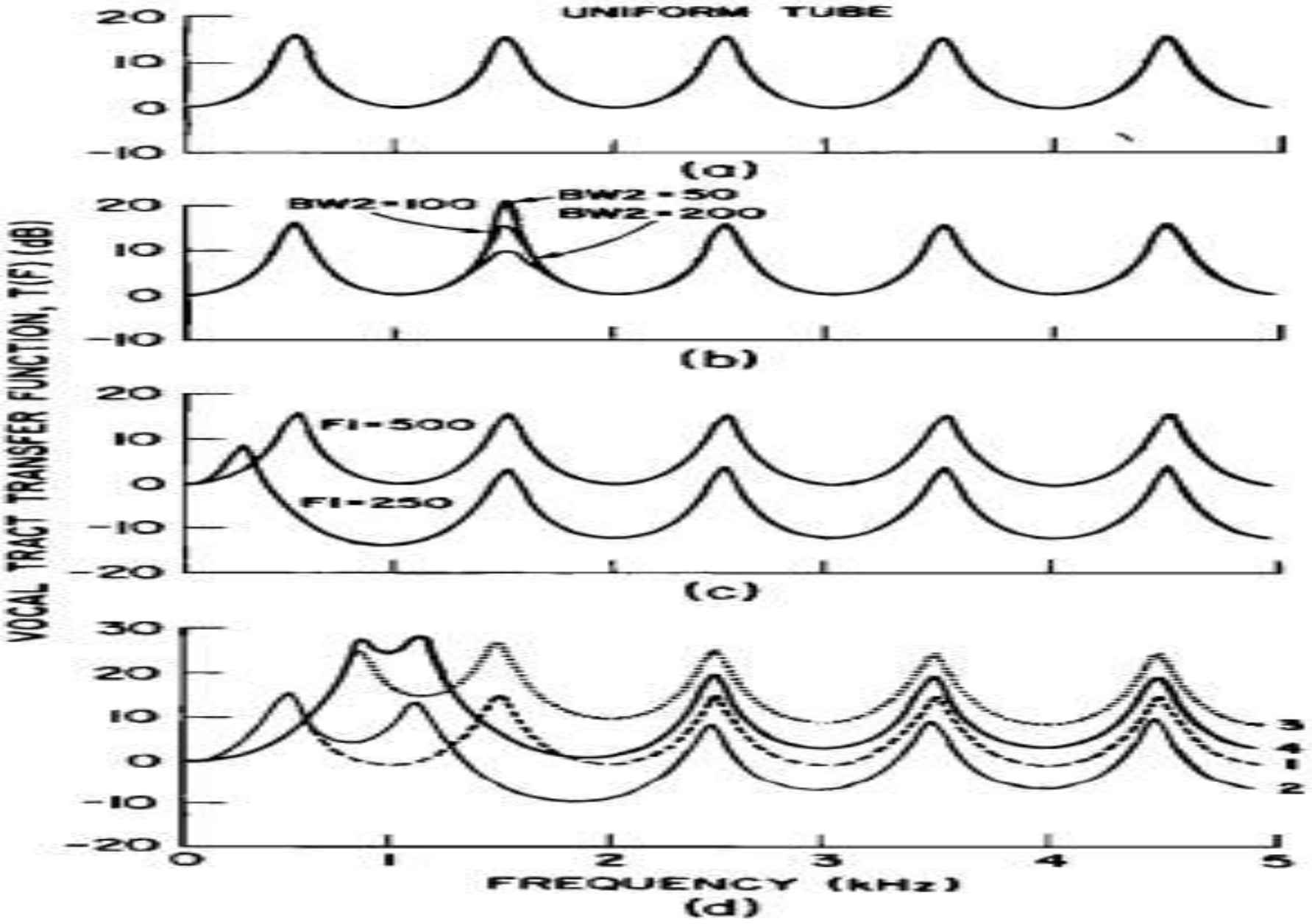
- Satisfactory approximation to the vocal tract transfer function for frication excitation can be achieved with a parallel set of digital formant resonators having amplitude controls, and no antiresonators
- Relatively simple rules for determination of the formant amplitude settings as a function of place of articulation can be derived from a quantal theory of speech production
- Only formants associated with the cavity in front of the oral constriction are strongly excited.

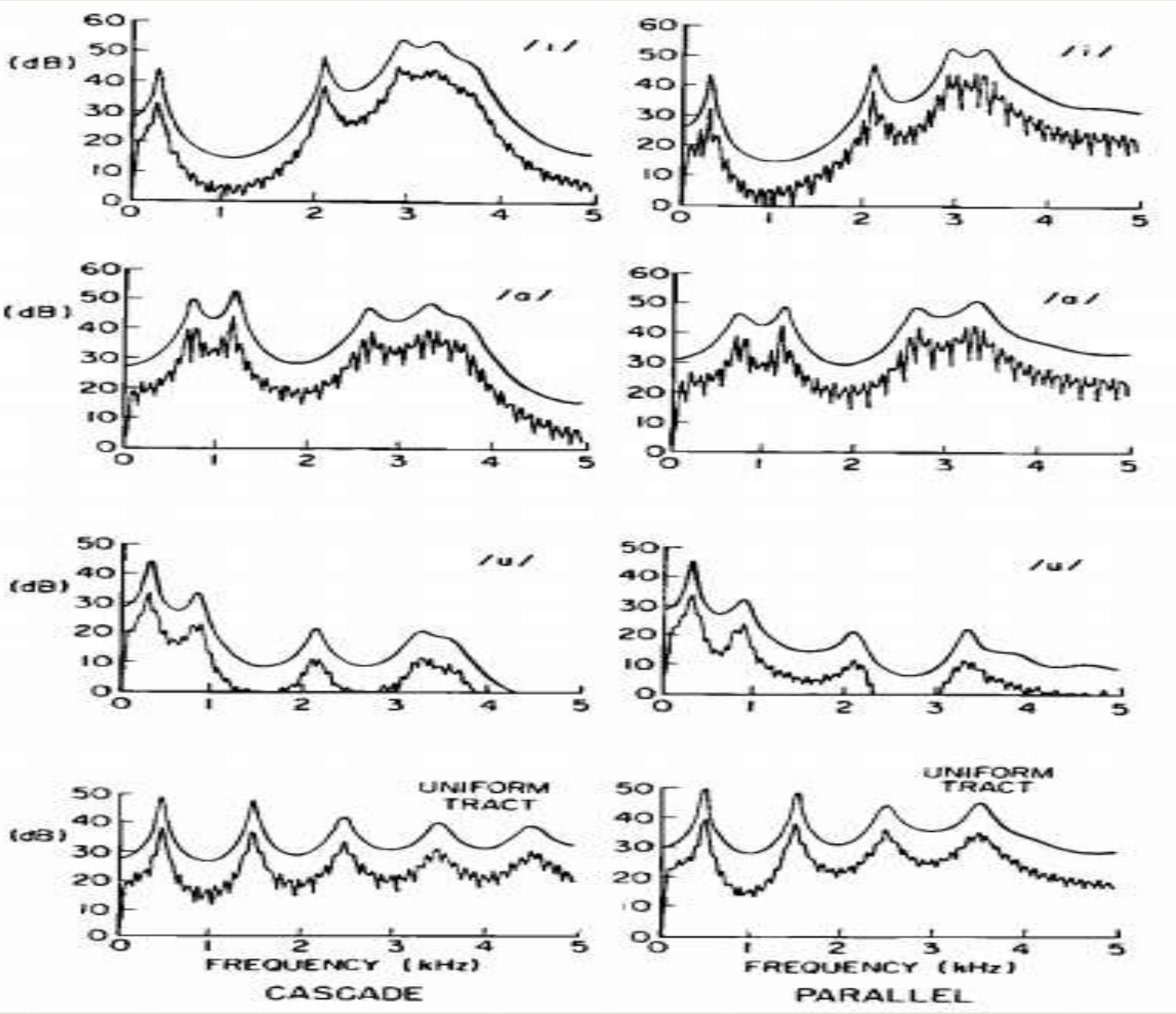
Contd…

- A sixth formant has been added to the parallel branch specifically for the synthesis of very high frequency noise in [s, z]
- During the production of a voiced fricative, output of the quasi-sinusoidal voicing source is sent through the cascade vocal tract model, while the frication source excites the parallel branch

Simulation of Cascade by Parallel Configuration

- The transfer function of the laryngealy excited vocal tract can also be approximated by five digital formant resonators connected in parallel
- What happens to formant amplitudes in the transfer function T(f) of a cascade model as the lowest five formant frequencies and bandwidths are change?


UNIFORM TUBE
(a)
BW2-100
BW2-50
BW2-200
(b)
F1-500
F1-250
(c)
VOCAL TRACT TRANSFER FUNCTION, T(F) (dB)
FREQUENCY (kHz)
(d)

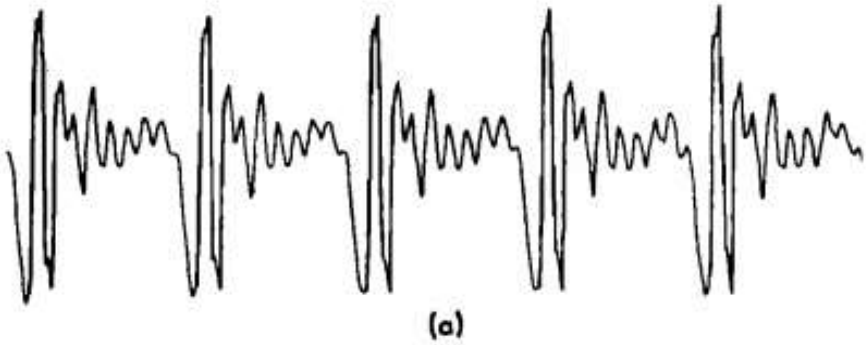

(dB)
/i/
/a/
/u/
UNIFORM TRACT
FREQUENCY (kHz)
CASCADE
PARALLEL

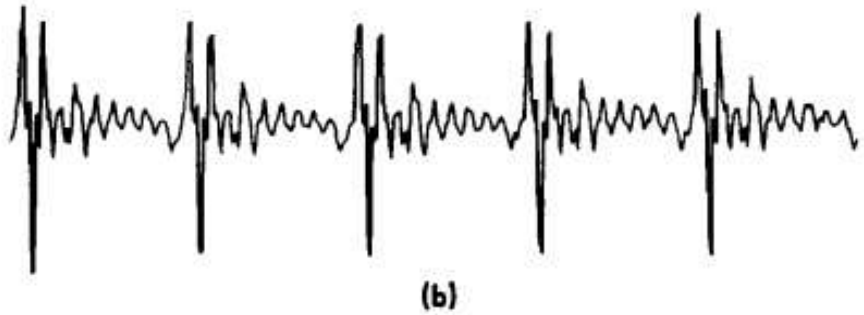
Radiation Characteristics

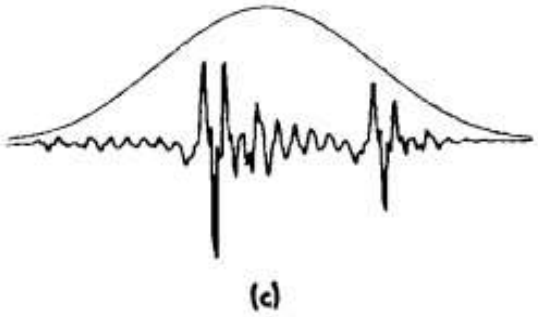
- The sound pressure measured directly in front of and about a meter from the lips is proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r, the distance from the lips
- The transformation is simulated in the synthesizer by taking the first difference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

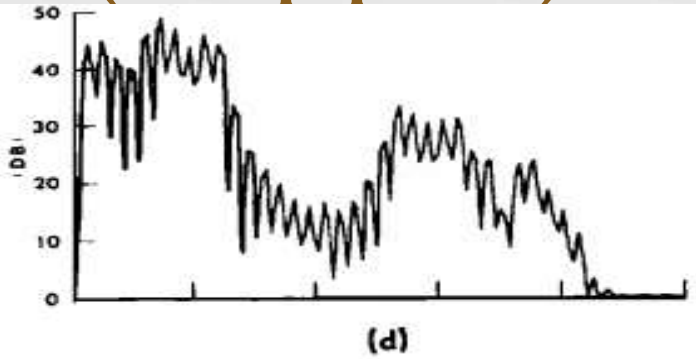
$$p(n) = u(n) - u(n-1)$$

Host Computer(skipped)









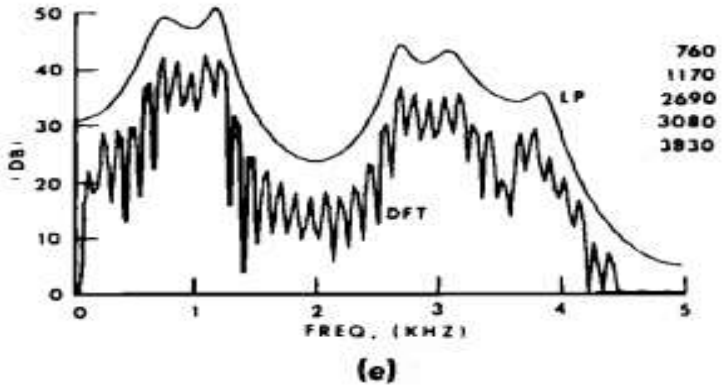


FIG. 16. The 25.6-ms (256 point) waveform segment extracted from a natural [a] with a fundamental frequency of 122 Hz, shown in (a) has been first differenced in (b), and multiplied by a Kaiser window in (c). The magnitude of the discrete Fourier transform of the non-preemphasized windowed waveform is shown in (d) and the DFT and magnitude of a linear prediction spectrum of the preemphasized windowed waveform is plotted in (e). Also listed are the frequencies of local maxima in the linear prediction spectrum; these maxima are usually good estimates of formant frequencies.

Vowel	F1	F2	F3	B1	B2	B3
[iʸ]	310	2020	2960	45	200	400
	290	2070	2960	60	200	400
[ɪᵊ]	400	1800	2570	50	100	140
	470	1600	2600	50	100	140
[eʸ]	480	1720	2520	70	100	200
	330	2020	2600	55	100	200
[ɛᵊ]	530	1680	2500	60	90	200
	620	1530	2530	60	90	200
[æᵊ]	620	1660	2430	70	150	320
	650	1490	2470	70	100	320
[ɑ]	700	1220	2600	130	70	160
[ɔᵊ]	600	990	2570	90	100	80
	630	1040	2600	90	100	80
[ʌ]	620	1220	2550	80	50	140
[oʷ]	540	1100	2300	80	70	70
	450	900	2300	80	70	70
[uᵊ]	450	1100	2350	80	100	80
	500	1180	2390	80	100	80
[uʷ]	350	1250	2200	65	110	140
	320	900	2200	65	110	140
[ɚ]	470	1270	1540	100	60	110
	420	1310	1540	100	60	110
[aʸ]	660	1200	2550	100	70	200
	400	1880	2500	70	100	200
[aʷ]	640	1230	2550	80	70	140
	420	940	2350	80	70	80
[oʸ]	550	960	2400	80	50	130
	360	1820	2450	60	50	160

- Syn
- mental
- st three
- lwidths
- effects
- nd falls
- ble
- timated

TABLE III. Parameter values for the synthesis of selected components of English consonants before front vowels (see text for source amplitude values).

Sonor	F1	F2	F3	B1	B2	B3
[w]	290	610	2150	50	80	60
[y]	260	2070	3020	40	250	500
[r]	310	1060	1380	70	100	120
[l]	310	1050	2880	50	100	280

Fric.	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
[f]	340	1100	2080	200	120	150	0	0	0	0	0	57
[v]	220	1100	2080	60	90	120	0	0	0	0	0	57
[θ]	320	1290	2540	200	90	200	0	0	0		28	48
[ð]	270	1290	2540	60	80	170	0	0	0	0	28	48
[s]	320	1390	2530	200	80	200	0	0	0	0	52	0
[z]	240	1390	2530	70	60	180	0	0	0	0	52	0
[š]	300	1840	2750	200	100	300	0	57	48	48	46	0
Affricate												
[č]	350	1800	2820	200	90	300	0	44	60	53	53	0
[ǰ]	260	1800	2820	60	80	270	0	44	60	53	53	0
Plosive												
[p]	400	1100	2150	300	150	220	0	0	0	0	0	63
[b]	200	1100	2150	60	110	130	0	0	0	0	0	63
[t]	400	1600	2600	300	120	250	0	30	45	57	63	0
[d]	200	1600	2600	60	100	170	0	47	60	62	60	0
[k]	300	1990	2850	250	160	330	0	53	43	45	45	0
[g]	200	1990	2850	60	150	280	0	53	43	45	45	0

Nasal	FNP	FNZ	F1	F2	F3	B1	B2	B3
[m]	270	450	480	1270	2130	40	200	200
[n]	270	450	480	1340	2470	40	300	300

Contd…

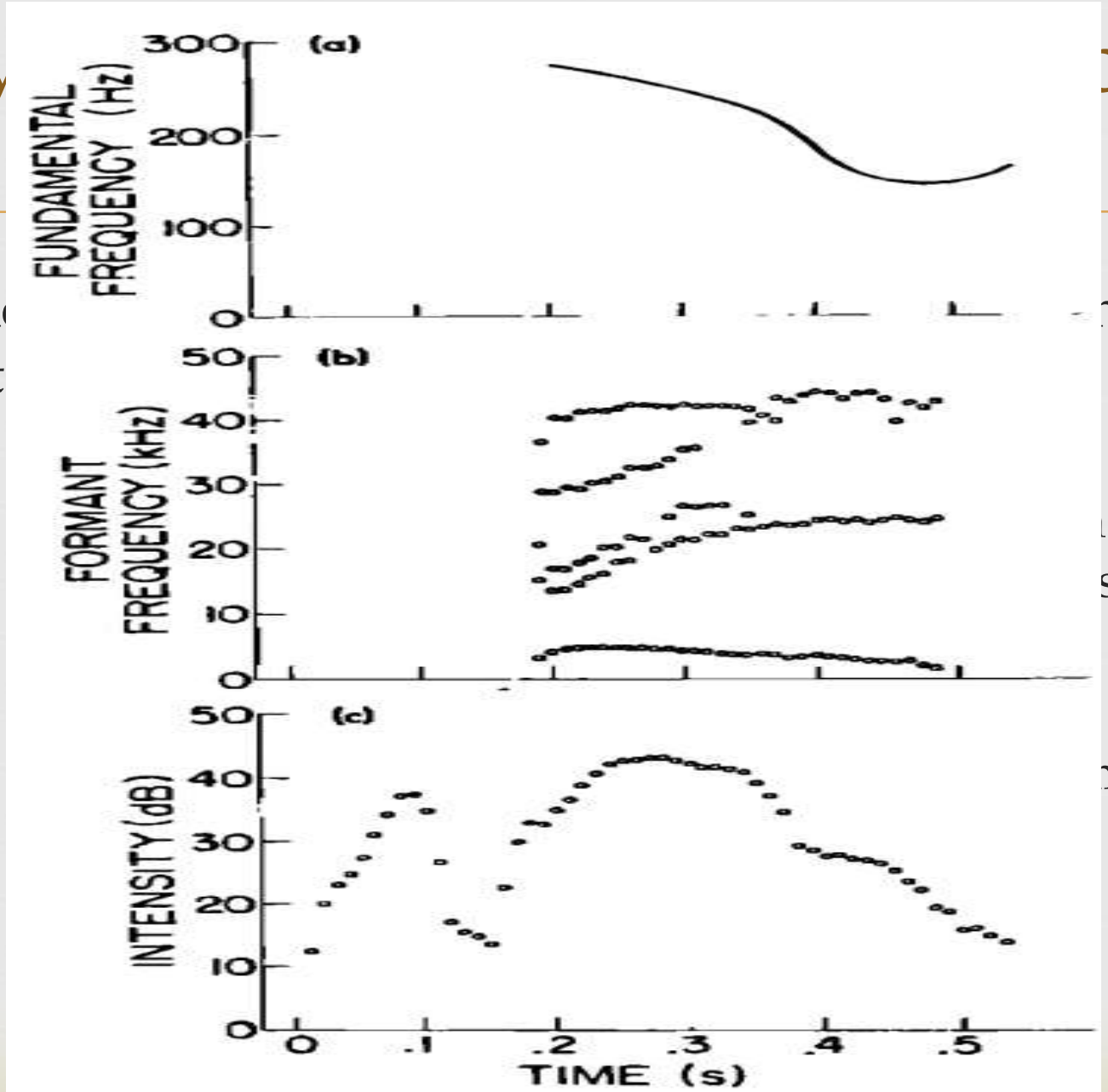
- The sonorant consonants are similar to vowels and require the same set of control parameters to be varied in order to differentiate among them
- AV, for a prevocalic sonorant should be about 10 dB less than in the vowel
- Voiceless fricatives (AF = 60, AV = 0, AVS = 0) and Voiced fricatives (AF= 50, AV= 47, AVS = 47)
- Formants to be excited by the frication noise source are determined by the amplitude controls A2, A3, A4, A5, A6, and AB
- Values presented in the table are appropriate only for consonants before front vowels

Contd…

- The parameters that are used to generate a nasal murmur include the nasal pole and zero frequencies FNP and FNZ
- A nasalized vowel is generated by increasing F1 by about 100 Hz, and by setting the frequency of the nasal zero to be the average of this new F1 value and 270 Hz

Sy ce

- The novel utt
  -
  - such
  - s and
  - close
  - m the



Conclusion

- The synthesizer is sufficiently flexible to generate good imitations to most if not all male and female voices
- It also appears possible to synthesize any phonetic sequence of English with excellent intelligibility
- A consonant-vowel synthesis cookbook that is based on this synthesizer is in preparation ☹