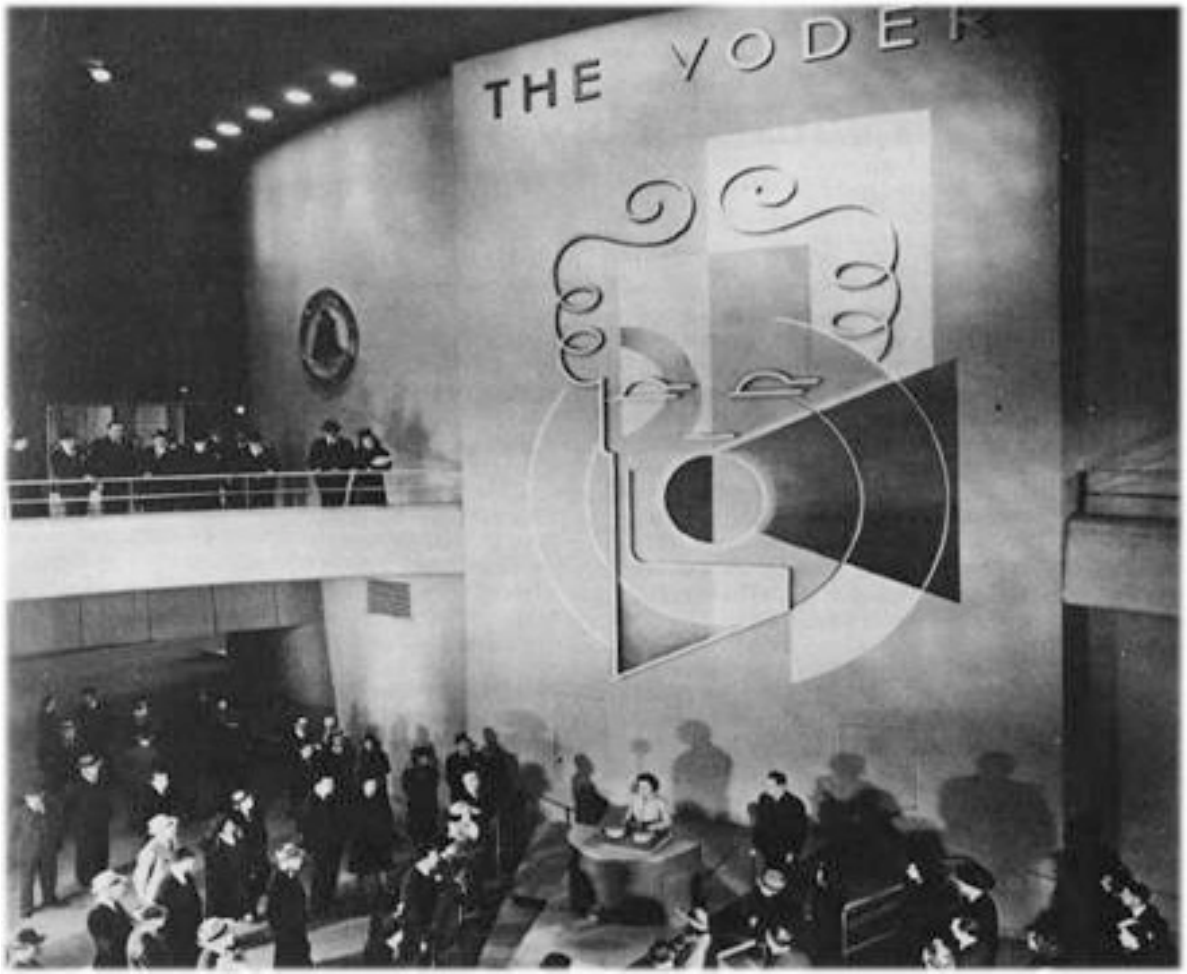
Why talking machines?

- Practical applications: text-to-speech
  - reading machines
  - query systems
  - adaptive technology for vocally challenged
- Theoretical interest
  - low-dimensional parameterization of speech signal
  - phonetic representation (copy-synthesis)
    - adequate to represent any utterance in any language
    - parameters can be identified that can be used for any speaker, any language (though of course values will differ).
  - automatic generation of parameters (rule synthesis)
    - explicit phonology-phonetics mapping

Early Synthesis

Homer Dudley's Voder (1939)

- Electronic circuits corresponding to formants
- Foot pedals for noise bursts
- Required manual operator



Early parametric synthesis

- Handcopying by painting through formants over spectrogram
  - The Pattern Playback designed by Franklin Cooper, 1951.
- Quantitative specification
  - Example 3 PAT, the "Parametric Articficial Talker" of Walter Lawrence, 1953.
  - Example 4 The "OVE" cascade formant synthesizer of Gunnar Fant
- Parallel vs. serial formant synthesizer: adequacy of copying
  - Example 5 Copying a natural sentence using Walter Lawerence's PAT formant synthesizer, 1962. (parallel)
  - Example 6 Copying the same sentence using the second generation of Gunar Fant's OVE cascade formant synthesizer, 1962. (cascade)
  - Example 7 Comparison of synthesis and a natural sentence, using OVE II, by John Holmes, 1961 (cascade)
  - Example 8 Comparison of synthesis and a natural sentence, John Holmes using his parallel formant synthesizer, 1973.(parallel)

Early parametric synthesis

- Female voices

  - Example 9 Attempting to scale the DECtalk male voice to make it sound female.
  - Example 10 Comparison of synthesis and a natural sentence, fremale voice, Dennis Klatt, 1986b,

- LPC analysis and re-synthesis

  - Example 13 Linear-prediction analysis and resynthesis of speech at a low-bit rate in the Texa Instruments Speak'n'Spell toy, Richard Wiggins, 1980.
  - Example 14 Comparison of synthesis and a natural recording, automatic analysis-resynthesi using multipulse linear prediction, Bishnu Atal, 1982.

Formant vs. LPC synthesis

- LPC copying can be done more automatically, as the LPC parameter extraction is analytical.
- Formant estimation is difficult and imperfect.
- But LPC parameters are statistical, values are specific to the speaker, context, sampling rate.
  - doesn't easily generalize to new speaker, context
  - How can they be used in automatic text-to-speech?

Automatic syllable-level (CV) synthesis

- Impossible to produce syllables by concatenation of C and V units (why?)
- Discovering rules
  - Example 15 Creation of a sentence from rules in the head of Pierre Delattre, using the Haskins Pattern Playback, 1959.
- Implementation of rules in computer code, generating parameters for formant synthesizer:
  - Example 16 Output from the first computer-based phonemic-synthesis-by-rule program, created by John Kelly and Louis Gerstman, 1961.
  - Example 17 Elegant rule program for British English by John Holmes, Ignatius Mattingly, and John Shearme, 1964.
- Implementation of rules for vocal-tract synthesizer:
  - Example 19 Rules to control a low-dimensionality articulatory model, by Cecil Coker, 1968.
- Concatenation of larger units

  Example 18 Formant synthesis using diphone concatenation, by Rex Dixon and David Maxey, 1968.

Klatt (1980): CV rules

- Tables of C,V parameters
- Interpolate values between adjacent segments
- Table values for Cs (stops especially) will be context-dependent.

Vowel parameters

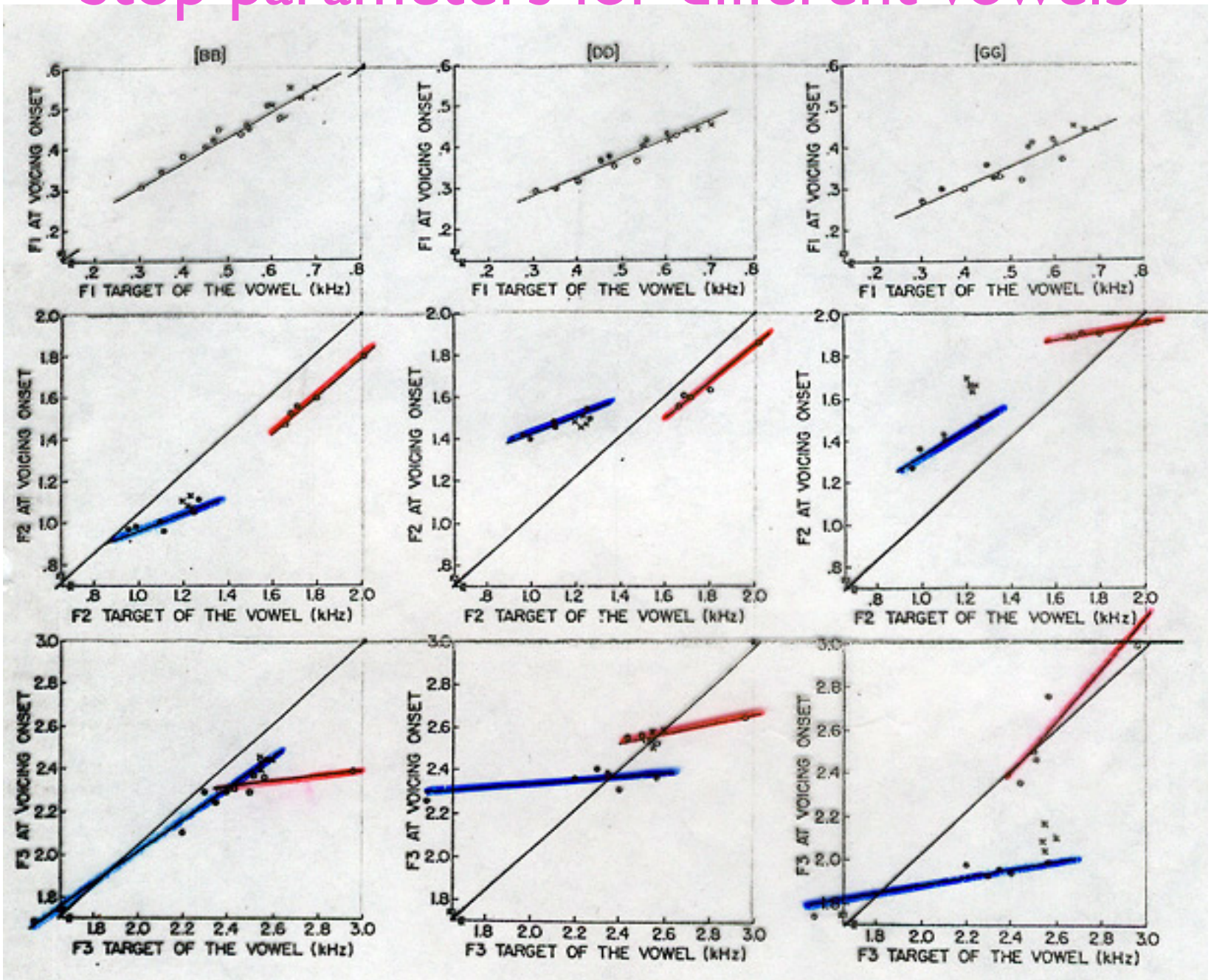
Vowel	*F*1	*F*2	*F*3	*B*1	*B*2	*B*3
[iʸ]	310	2020	2960	45	200	400
	290	2070	2960	60	200	400
[Iᵊ]	400	1800	2570	50	100	140
	470	1600	2600	50	100	140
[eʸ]	480	1720	2520	70	100	200
	330	2020	2600	55	100	200
[ɛᵊ]	530	1680	2500	60	90	200
	620	1530	2530	60	90	200
[æᵊ]	620	1660	2430	70	150	320
	650	1490	2470	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

Parameters for consonants in front vowel contexts

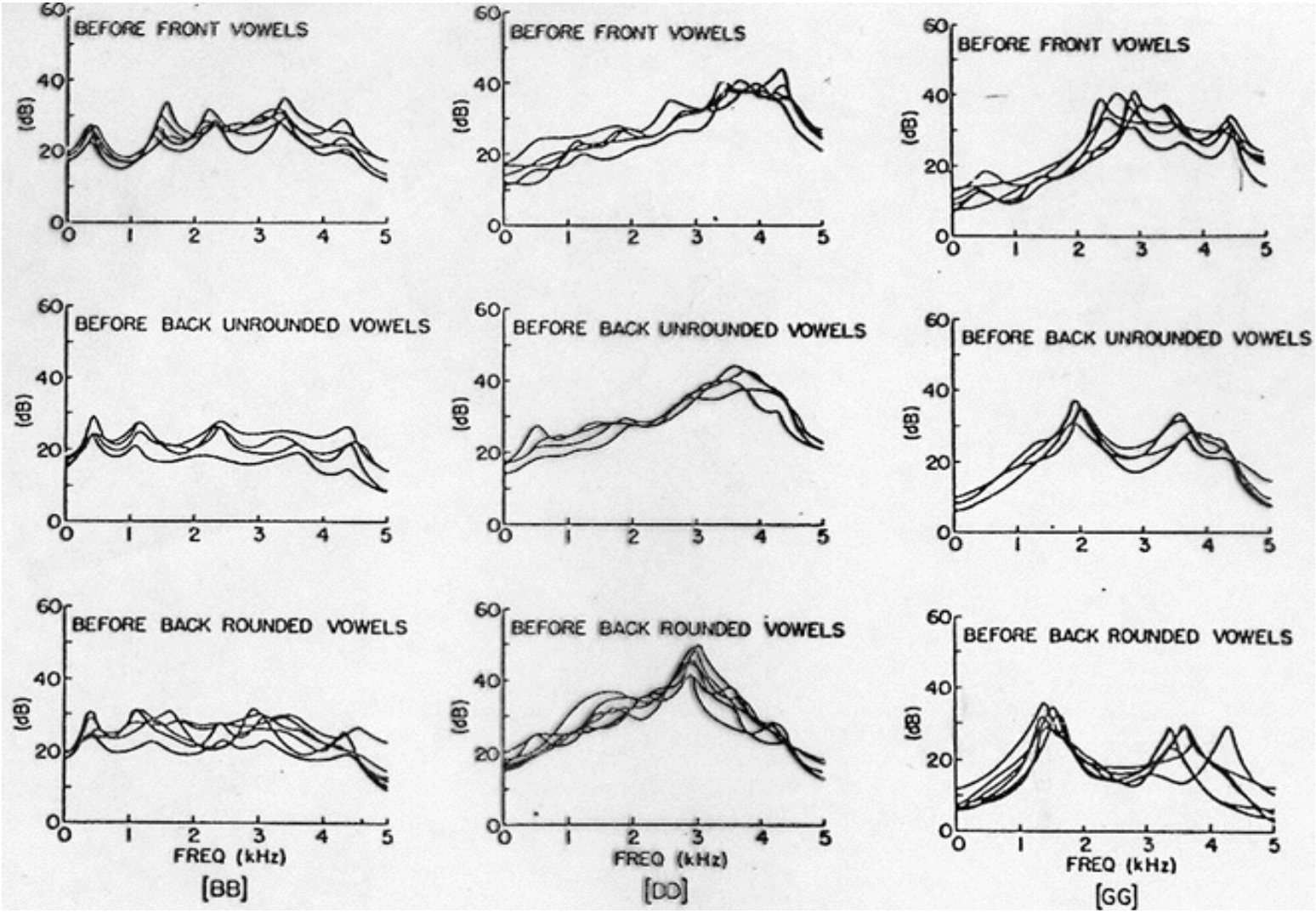
TABLE III. Parameter values for the synthesis of selected components of English consonants before front vowels (see text for source amplitude values).

Sonor	*F*1	*F*2	*F*3	*B*1	*B*2	*B*3						
[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.	***F*1**	***F*2**	***F*3**	***B*1**	***B*2**	***B*3**	***A*2**	***A*3**	***A*4**	***A*5**	***A*6**	**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

Stop parameters for different vowels



Stop burst spectra for different vowels



Sentence synthesis-by-rule

- Requires explicit knowledge of phonological representation and rules in a language
  - prosodic rules
  - alternation of C,V units in context

- Rules for formant synthesizer:
  - Example 20 First prosodic synthesis by rule, by Ignatius Mattingly, 1968.
  - Example 21 Sentence-level phonology incorporated in rules by Dennis Klatt, 1976.

- Rules for concatenation of lpc-coded units:
  - Example 22 Concatenation of linear-prediction diphones, by Joe Olive, 1977.
  - Example 23 Concatenation of linear-prediction demisylables by Catherine Browman, 1980.

Complete text-to-speech systems

Example 24 The first full text-to-speech system, done in Japan by Noriko Umeda et al., 1968.

- Example 25 The first Bell Laboratories text-to-speech system by Cecil Coker, Noriko Umeda, and Catherine Browman, 1973.
- Example 26 The Haskins Laboratories text-to-speech system, 1973.
- Example 30 The M.I.T. MITalk system by Jonathan Allen, Sheri Hunnicut, and Dennis Klatt, 1979.
- Example 33 The Klattalk system by Dennis Klatt of M.I.T. which formed the basis for Digital Equiptment Corporation's DECtalk commercial systenm 1983.
- Example 34 The AT&T Bell Laboratories text-to-speech system, 1985.

Recent Approaches

- Gestural models
  - rules not required, as context-sensitive acoustic properties emerge from:
  - gestural overlap
  - reduction in activation
- Unit selection (concatenation)
  - lpc-coded corpus is segmented automatically (forced alignment) at many levels (phone, diphone, syllable, word)
  - choice of unit size determined by match between units stored in corpus and units required for the utterance to be synthesized.
- HMM (Hidden Markov Model)
  - speech is represented as a probabilistic sequence ("chain") of "hidden" states, each of which is associated with different probabilities of observed acoustic events.