REVIEW OF TEXT-TO-SPEECH CONVERSION FOR ENGLISH

Presented by
Najeeb Khan
2013-3-15

INTRODUCTION

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

Last Presentation

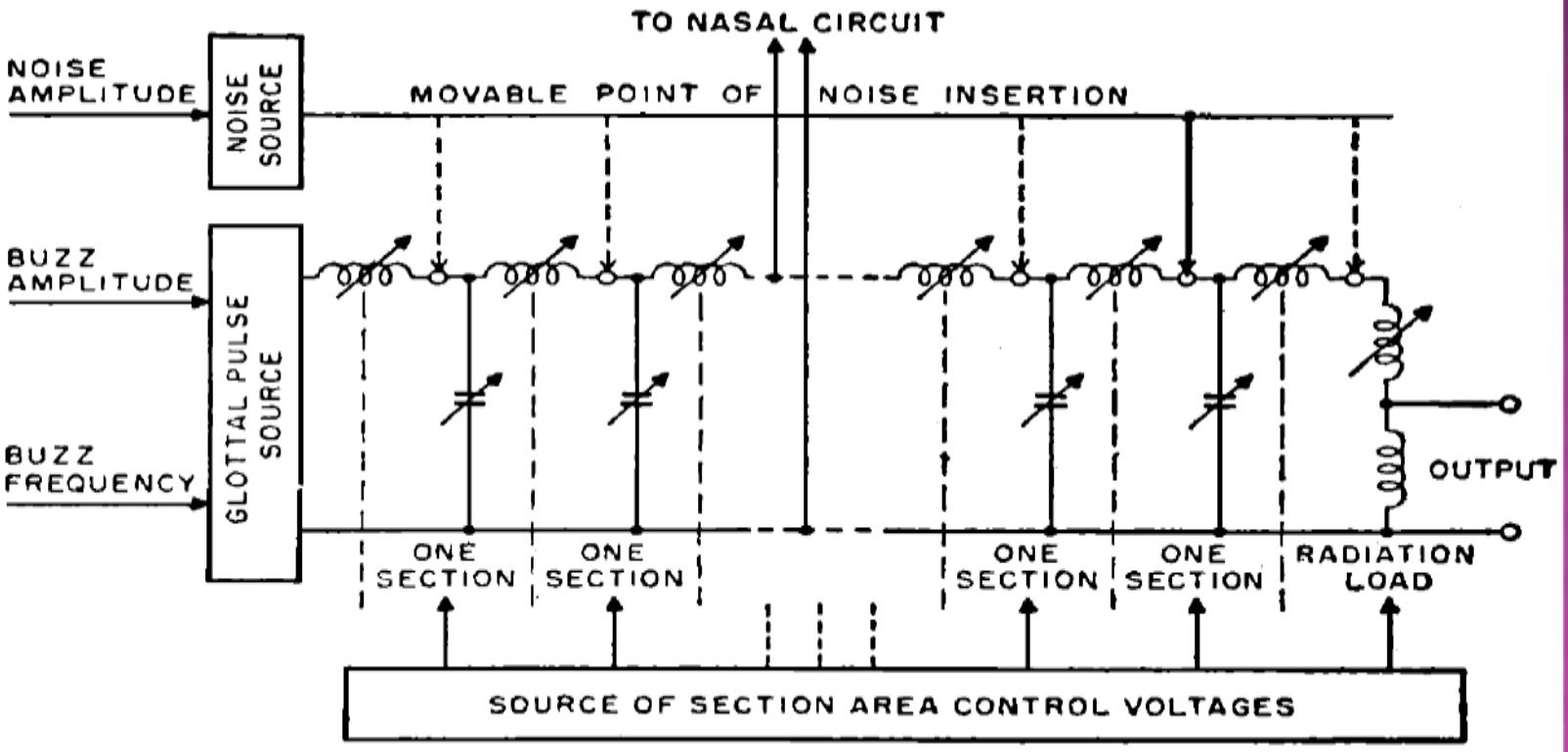
- Modern formant synthesizers of several different configuration are capable of imitating male speakers nearly perfectly.
- Formant synthesizers results in unsatisfactory imitation of breathy high-pitched vowels.

ARTICULATORY MODELS

- In an articulatory model the tube corresponding to the vocal tract is usually divided into many small sections, and each section is approximated by an electrical transmission line analog

CONTD…

- Stevens 1953: First Electronic model
- Rosen 1958: Dynamic Analog of vocal tract



CONTD…

- Hecko 1962: Added side branch to approximate nasal tract.
- Modern Improved simulations focused on
  - Frequency dependent loss terms
  - Provision for cavity wall motions
  - Time varying termination impedance of glottis

CONTD...

- First articulatory synthesizer used sawtooth current source
- Since volume velocity depend very little on the shape or impedance of the vocal tract at least for vowels.
- Research focus: better approximation to the vibration pattern and resulting volume velocity waveform.

MECHANICAL MODELS OF GLOTTIS

- Flanagan 1968: Mass spring damping system
  - Waveforms generated by this model bore many similarities to physiological waveforms.
- In natural vibrations the upper and lower surface of the folds vibrate out of phase.
- The first order aspects of this phenomenon have been captured by two mass models of each fold, in which the upper and lower surfaces of the folds are simulated by separate masses coupled by a spring.

CONTD…

- Flanagan 1975: Such a model coupled to a digital simulation of a transmission line analog of vocal tract were demonstrated.
- Titze 1974: Modeled vocal fold vibration by a three dimensional structure of a number of coupled masses.
- No entirely satisfactory solution has been proposed for simulating what happens when the vocal folds slam together at the midline and deform in some way to absorb the energy of the impact.

SOURCE VOCAL TRACT INTERACTION

- The resonance structure of the vocal tract result in standing pressure waves that can have an effect on the pressure distribution at the glottis and hence the vibration pattern and airflow waveform from the voicing source.
- The vibrating glottis provide a time varying termination impedance that affects the formants and bandwidths of the vocal tract transfer function.

CONTD...

- These effects are not large but may be of some importance in simulating natural voice quality.
- Liljencrants(1985) has programmed a detailed articulatory model to simulate these effects with the results that synthesis of steady vowels sound quite natural.

PROBLEMS IN ARTICULATORY MODEL

- Precise acoustic aspects of complex articulatory model are not known.
- Greater computational cost of the articulatory synthesis precludes its use at present time

AUTO ANALYSIS/RESYNTHESIS OF NATURAL WAVEFORMS

- Linear prediction
  - At least in the absence of source excitation, the next sample of a speech waveform can be estimated from a weighted sum of 10 or so previous waveform samples, the weights being the linear prediction coefficients.
- If the source waveform can be found by other means and if predictor coefficients are updated every 10ms, a good approximation to the original waveform can be derived from this low bit rate representation.

LINEAR PREDICTION FOR TTS

- A problem arises when going from duplicating a natural utterance to the more difficult task of creating new sentences by concatenating pieces of speech.
- The predictor equations do not estimate formant frequencies and bandwidth accurately.
- For same f_0 its not a problem, but for synthesis f_0 the first formant may be in error by 8% or more and BW can be seriously deviant.

CONTD...

- Other analysis by synthesis procedures:
  - It has even been possible to mimic a high pitched female singing voice by summing together, for each period, formant like damped sinusoid waveforms.
- But the problem is of preserving naturalness in general text to speech application.

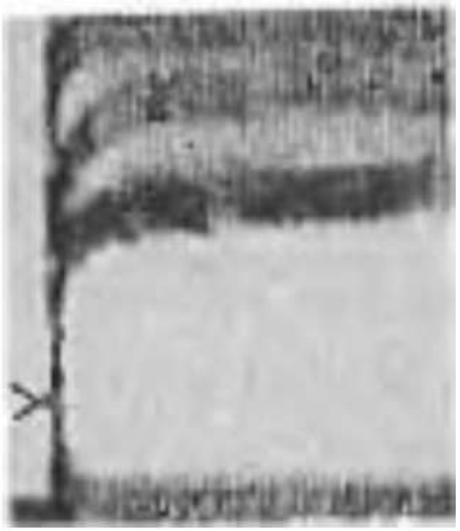
CONCLUSION

- Linear prediction is a powerful method of duplicating an utterance with high fidelity, but there are limitations on its applicability to general text synthesis.
- An articulatory model will be the ultimate solution but computational costs and lack of data upon which to base rules prevent immediate application of this approach.

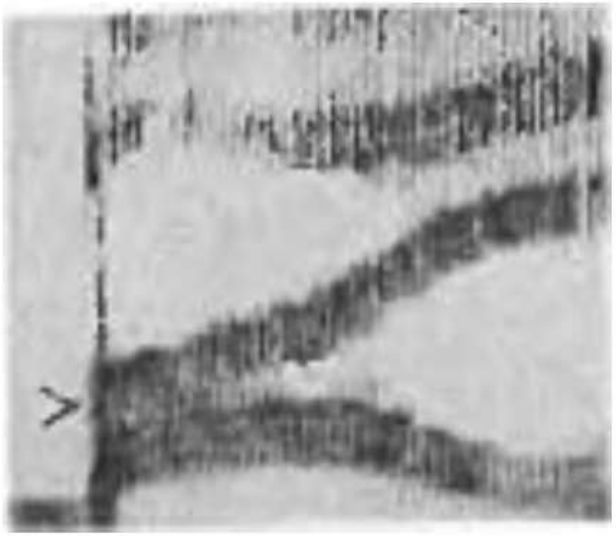
ACOUSTIC PROPERTIES OF PHONETIC SEGMENTS

- To generate speech using a formant synthesizer it is necessary to develop rules to convert sequences of discrete phonetic segments to time varying control parameters.
- Such rules depend on data obtained by acoustic analysis of speech.
- Perceptual data establishing the sufficiency of individual acoustic cues are also of considerable value in determining a rule strategy.

CONTD...



bi (*bee*)



baɪ (*buy*)

bou (*bow*)

SOUND SPECTROGRAPH

- The investigation of acoustic cues having the greatest importance for different speech sounds began with the use of the sound spectrograph machine.
- Broadband Spectrogram
  - A plot of frequency versus time in which darkness represents the energy present within a 300Hz bandwidth as averaged over about 2-3ms.
- The display is designed to represent formants as slowly varying horizontal dark bands and to indicate f_0 as the inverse of the temporal spacing between vertical striations

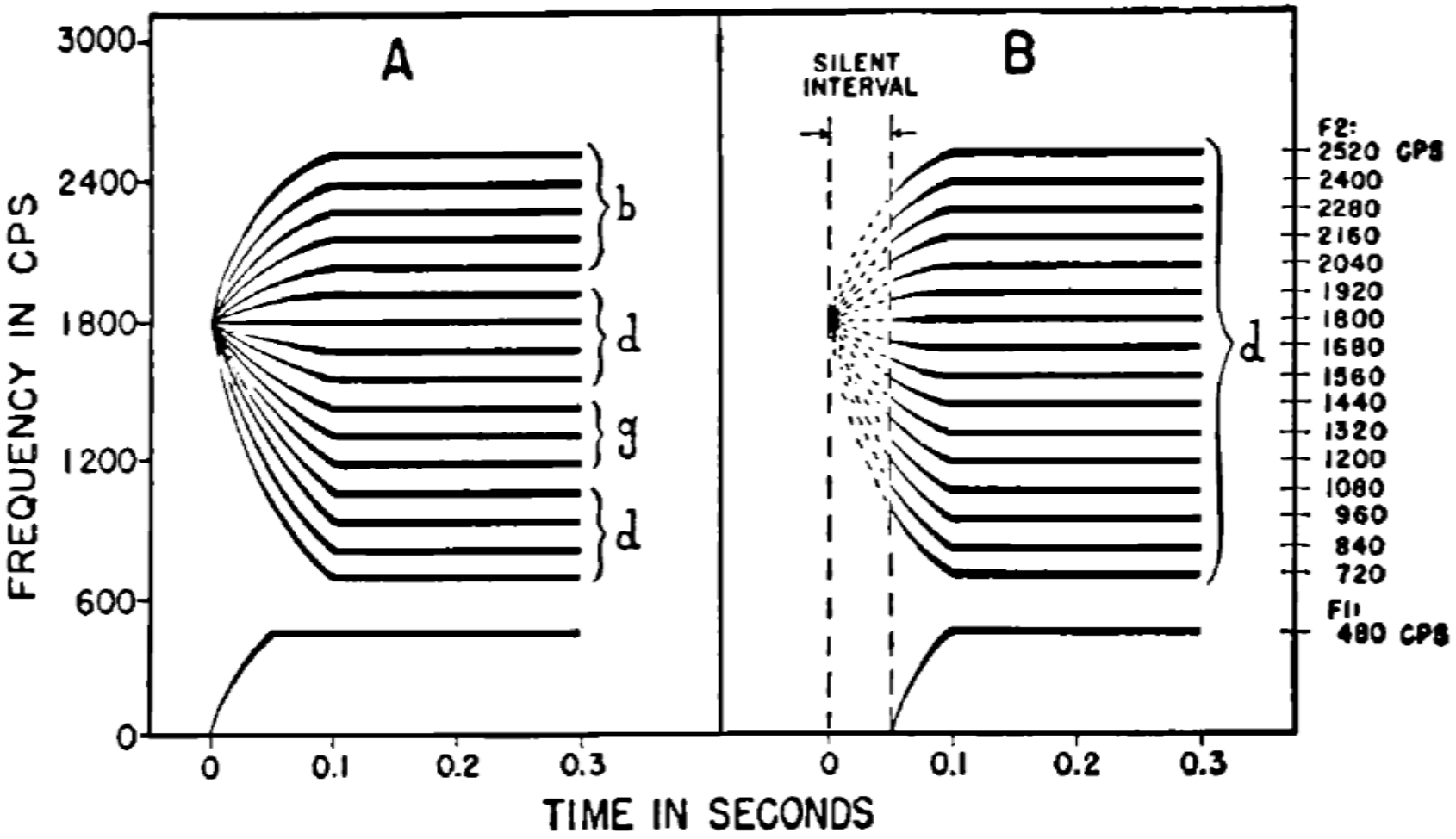
SYNTHESIS BY ART

- Cooper created stylized versions of syllables in an effort to determine the acoustic cues sufficient for the synthesis of slected phonetic contrasts. This resulted in explicit rules for synthesis of English speech sounds.
- Research suggested the importance of formant frequency, formant frequency motions, spectral peaks in noise, relative timing onsets in different frequency regions as cues for voicing, manner and place of articulation.

ENCODED NATURE OF SPEECH

- 1967: The researchers emphasized the encoded nature of speech
  - Acoustic cues to the identity of a phoneme were spread out in time so as to overlap with cues for adjacent phonemes.
  - Cues were context dependent, e.g. the same plosive burst was heard as a different consonant depending on the vowel that followed.
  - There appeared to be no one invariant acoustic cue signaling the presence of a given stop consonant, rather the consonant identity have to be inferred from formant transitions into adjacent vowels.

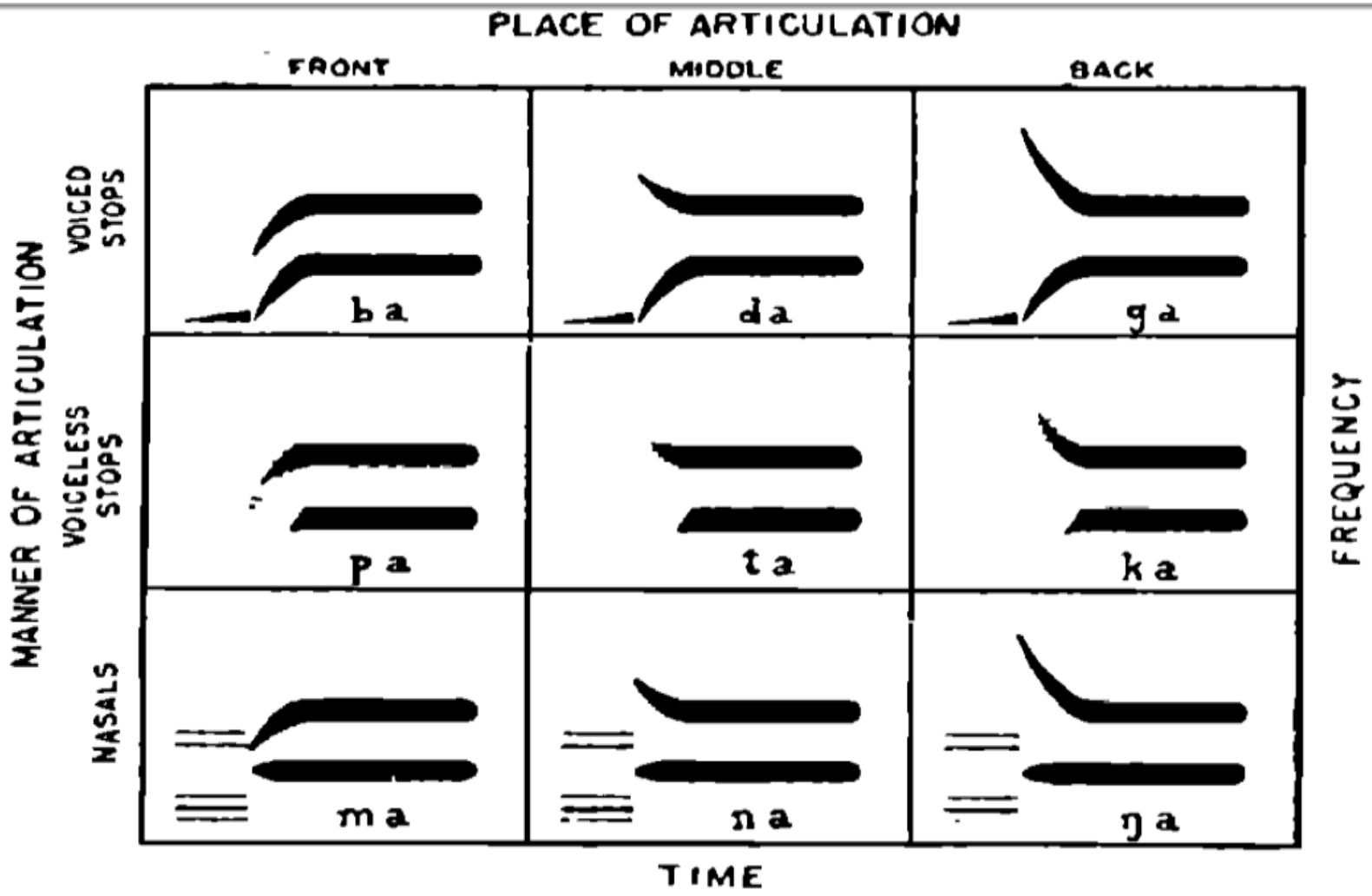
LOCUS THEORY (DALATTRE)



CONTD…

- Dalattre found that if the second formant actually started at 1800Hz in each case rather than at various values shown in the figure, listeners heard /bi, da, gu/ instead of /di, da, du/
- /g/ required two loci for front and back vowels.

CONTD...


PLACE OF ARTICULATION
FRONT
MIDDLE
BACK
MANNER OF ARTICULATION
VOICED STOPS
VOICELESS STOPS
NASALS
ba
da
ga
pa
ta
ka
ma
na
ŋa
FREQUENCY
TIME

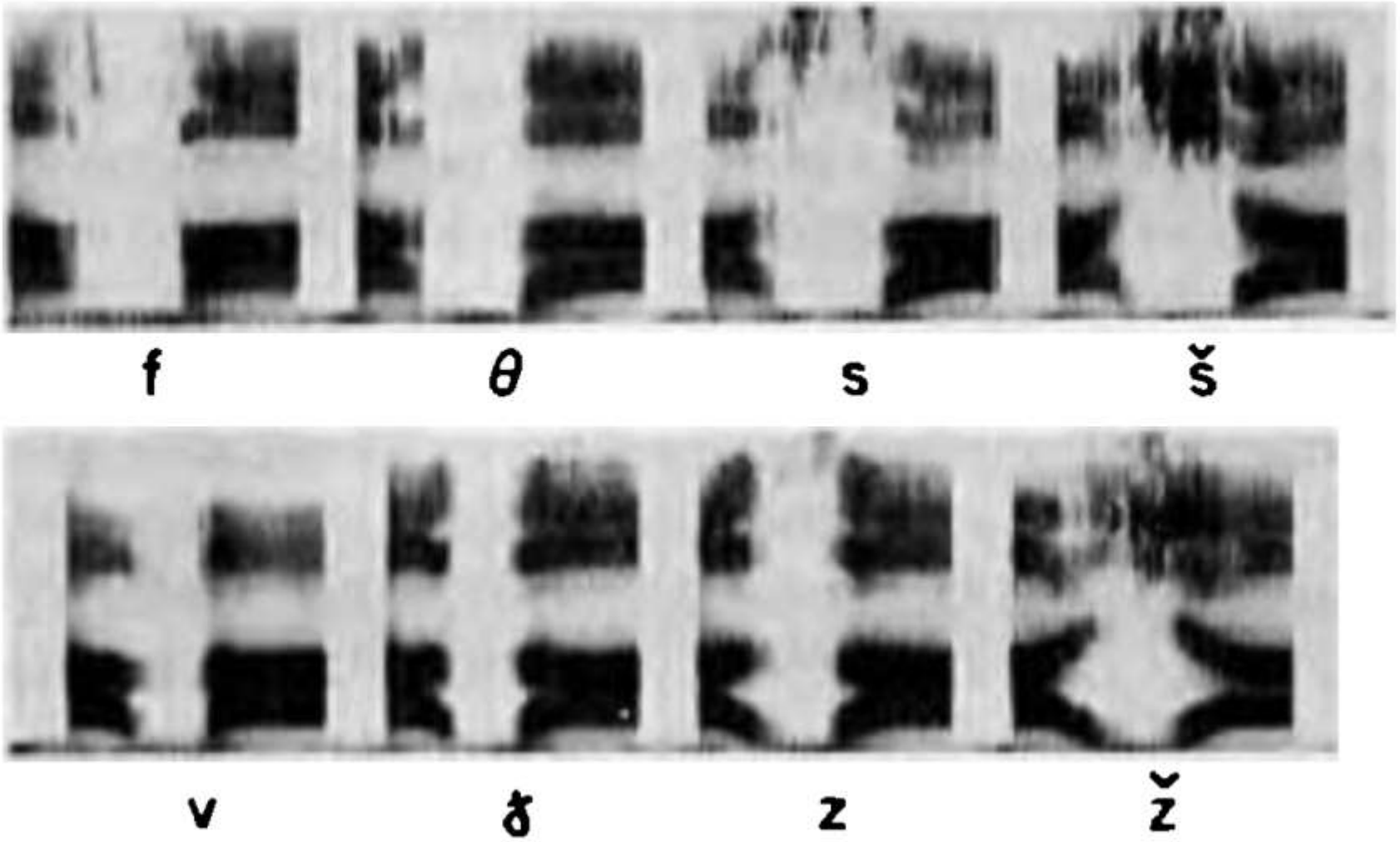
VOWELS

- Vowels can be represented by an all pole vocal tract transfer function.
- Relative magnitude of formant peaks can be predicted from formant frequencies.
- English vowels can be described in terms of
  - The frequencies of lowest 3 formants
  - Any formant motions associated with diphthongization
  - Differences in vowel duration.
- Formant bandwidth also differ slightly among vowels.

SONORANT CONSONANTS

and 12.

FRICATIVES



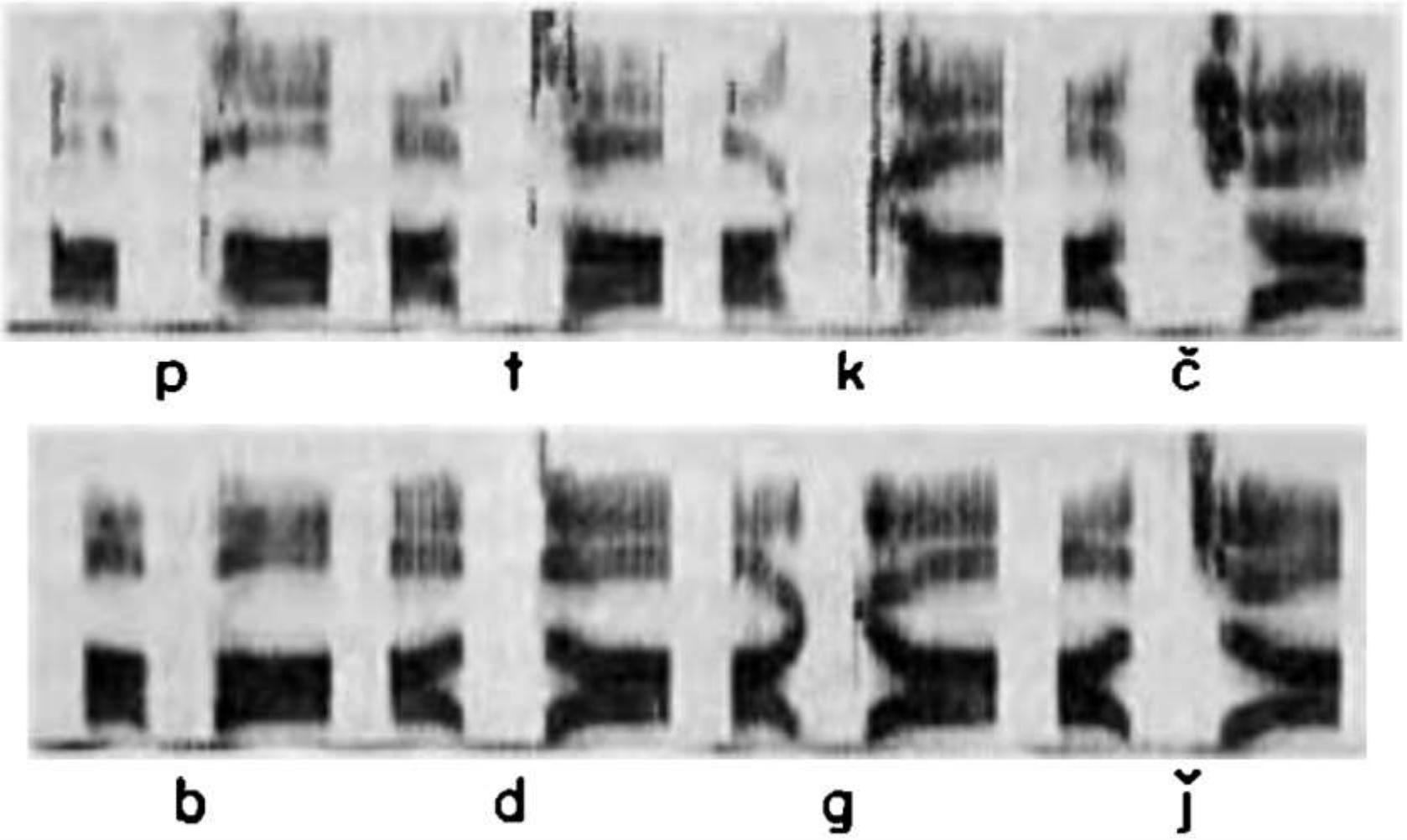
CONTD...

- Each fricative noise has a relative fixed characteristic shape.
- However lip rounding for a rounded vowel may lower the frequencies of the most prominent spectral peaks slightly.
- Distinction between /f/ and /Θ/ depends on the vocal tract shape of adjacent vowels.

PLOSIVES

- The voiced plosives /b,d,g/ consists of a closure interval, a brief burst of noise at release and formant transitions into and out of adjacent segments.
- /b,d,g/ include evidence of voicing during closure i.e periodic low frequency energy known as voice bar.

CONTD...



NASALS

- The nasal consonants /m,n,N/ consists of a murmur during the interval when the oral cavity is closed and rapid transition into and out of adjacent segments.

SEGMENTAL SYNTHESIS BY RULE PROGRAMS

- A synthesis by rule program constitutes a set of rules for generating what are often highly stylized and simplified approximations to natural speech.
- The rules consider only the cues which are important for each phonetic contrast.

SPEECH SYNTHESIS TECHNIQUES

- Heuristic rules to control formant synthesizer
- Articulatory rules to control model of the vocal tract.
- Strategies for concatenating pieces of encoded natural speech.

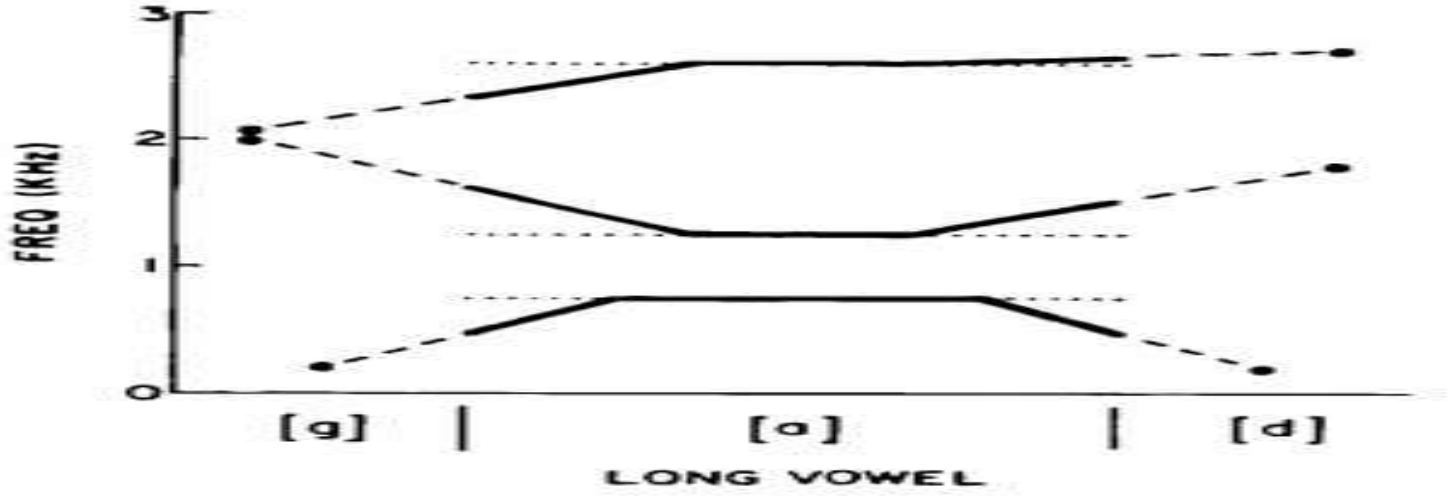
FORMANT BASED RULE PROGRAMS

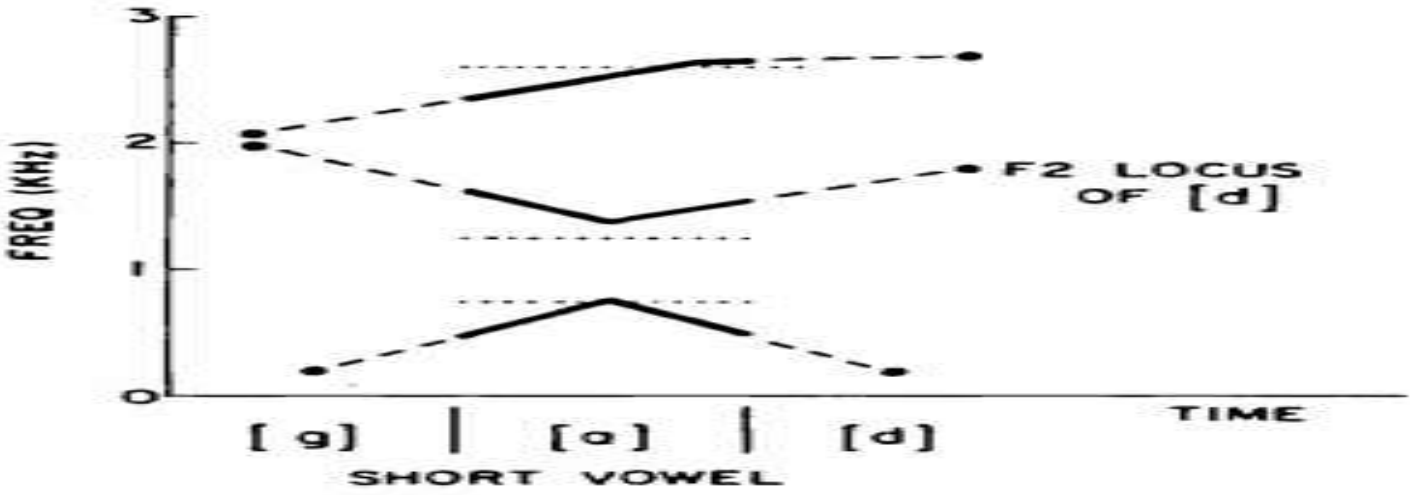
- Kelly and Gerstman(1961, 1964): First synthesis by rule program capable of synthesizing speech from phonetic representation.
- It used cascaded three formant synthesizer excited by impulse train/noise.
- Duration and f_0 contour were copied from natural speech.

HOLMES (BRITISH PHONEMES)

- Parallel formant synthesizer
- Simple parameter generation algorithm, whose operation was determined entirely by values in tables.

CONTD...


3
2
1
0
FREQ (KHz)
[g]
[o]
[d]
LONG VOWEL


3
2
1
0
FREQ (KHz)
F2 LOCUS
OF [d]
TIME
[g]
[o]
[d]
SHORT VOWEL

MATTINGLY (AMERICAN PHONEMES)

- Formant transitions that were s-shaped rather than linear.
- Used set of letter to sound rules and a 140,000 word phonemic dictionary.
- Experimental Haskins TTS system for reading machine.
- Terminated because of funding lapse.

RABINAR SYNTHESIZER

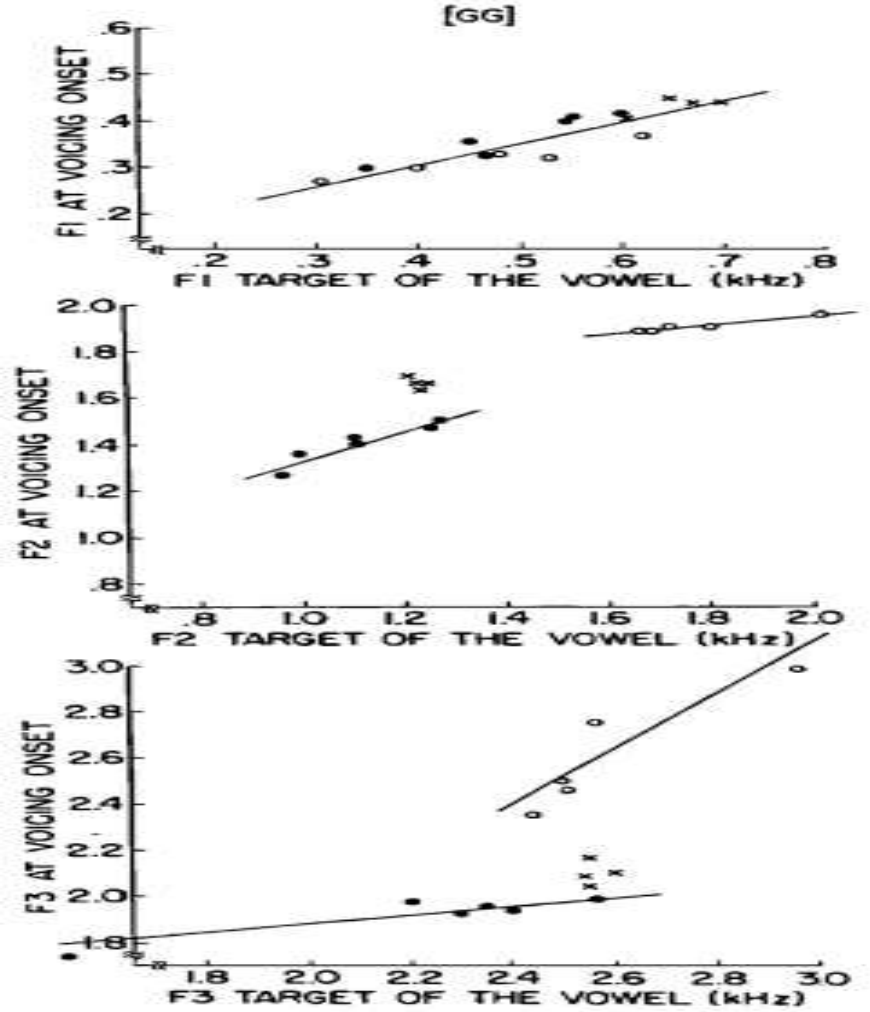
- Used a critically damped second order smoothing filter to constrain formant frequencies to move constantly in time as required by acoustic theory.
- Synthesized CV and VC nonesense syllables with consonantal intiligibility of about 75%

KLATT

- Extended the earlier work by formulating rules for generating CVC syllables with greater fidelity to measured characteristics of English consonants.
- Hybrid cascade/parallel synthesizer and rule programs that allowed specification of targets and straight line transitions.
- Achieved 95% intelligibility for CVC nonesense syllables.

CONTD…

- F_{2onset}
- Divid nto front roundc



ARTICULATION BASED RULE PROGRAMS

- Several research groups attempted to devise simplified models of the articulators or models of the observed shapes of the vocal tract.
- Kelly and Lockbaum: Stored tables of area functions for each phonetic segment and a linear interpolation scheme.
- Latter models abandoned the direct specification of an area function in favor of an intermediate model possessing a small set of movable structures corresponding to tongue, jaw, lips, velum and larynx.
- Such models had control problems.

CONTD...

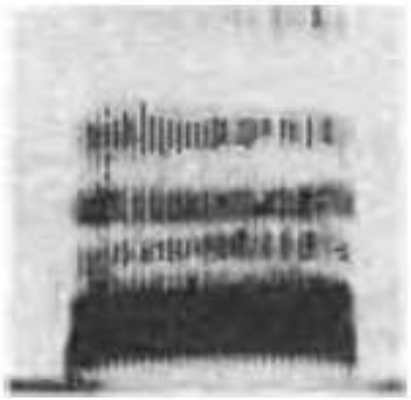
- An entire TTS system based on an articulatory model was created in japan.
- The text analysis and pause assignment rules of this system were based on a sophisticated parser.

RULE COMPILER

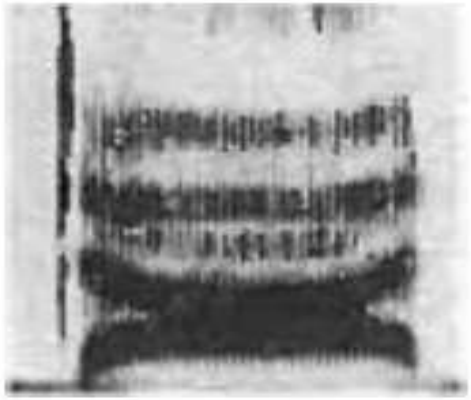
- Special programming language to permit linguists to formulate synthesis rules in a natural way.
- An important advantage of the language is to refer to natural sets of phonemes through a distinctive feature notation, making rule statement simple, efficient and easy to read.
- Some rules operate on phonetic segments while others on syllables or whole words and phrases
- Rule programs distinguish between level of description.

CONCATENATION SYSTEMS

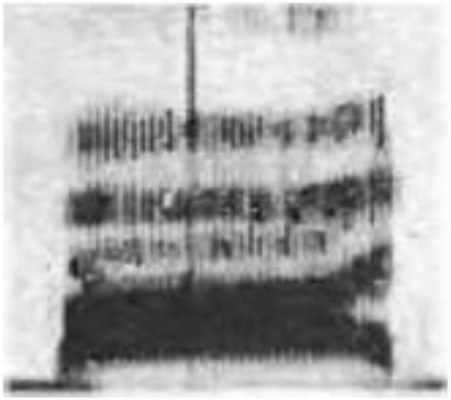
- Pieces of natural speech are used as building blocks to reconstitute an arbitrary utterance.
- Phoneme concatenation has failed because of the co-articulation
- Co-articulatory influences tend to be minimal at the acoustic center of the phoneme
- Thus the DIPHONE, the acoustic chunk from the middle of one phoneme to the middle of the next phoneme is an attractive unit.



[bab]

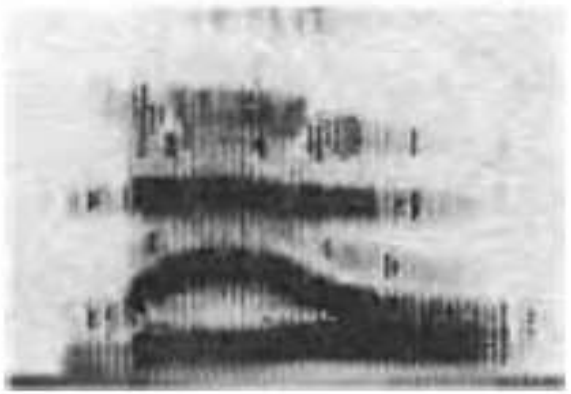


[dad]

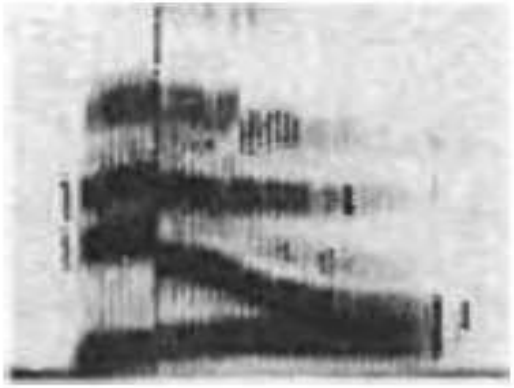


[bad]

[brb]



[ɻrɻ]



[brɻ]

CONTD...

- A closely related alternative to diphone is the demi-syllable i.e half of a syllable.
- About 1000 demi-syllables
- Highly co-articulated syllable-internal consonants are treated as units.
- Co-articulation across the syllables is not handled well