DOCUMENT RESUME

ED 028 432 48 AL 001 834

By-Vanderslice, Ralph; Rand, Timothy

Synthetic Intonation.

Michigan Univ., Ann Arbor. Center for Research on Language and Language Behavior.

Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research.

Bureau No-BR-6-1784

Pub Date 1 Feb 69

Contract-OEC-3-6-061784-0508

Note-22p.; Report included in Studies in Language and Language Behavior, Progress Report No. VIII.

EDRS Price MF-$0.25 HC-$1.20

Descriptors-Algorithms, *Artificial Speech, *Computational Linguistics, Input Output, *Intonation

Identifiers-Discourse Synthesis, *Synthetic Intonation

Pitch-synchronous, time-domain operation on digitized waveforms of human speech produces artificial changes in prosodic parameters, especially fundamental frequency and rhythm. Pitch of voiced segments is raised or lowered by an algorithm which truncates or "pads," respectively, each pitch period in the stored vector by an appropriate amount. Durations are altered by reduplicating or deleting pitch periods as necessary. Speech output, though of telephone quality, is more natural and intelligible than most fully synthetic speech. Potential applications are varied and far-reaching. (Author/DO)

BR-6-1784
PA-48

OE-BR



# SYNTHETIC INTONATION[1]

Ralph Vanderslice and Timothy Rand

Center for Research on Language and Language Behavior
The University of Michigan

Pitch-synchronous, time-domain operation on digitized waveforms of human speech produce artificial changes in prosodic parameters, especially fundamental frequency and rhythm. Pitch of voiced segments is raised or lowered by an algorithm which truncates or "pads", respectively, each pitch period in the stored vector by an appropriate amount. Durations are altered by reduplicating or deleting pitch periods as necessary. Speech output, though of telephone quality, is more natural and intelligible than most fully synthetic speech. Potential applications are varied and far-reaching.

Speech synthesizers, especially formant-resonator terminal analogs, have proven exceedingly useful for testing various models of natural language production and perception. The speech synthesizer constitutes a sort of "vocal Übermarionette" which performs exactly and only as it is explicitly programmed. This is of course its great virtue for speech research: unlike any human talker, the synthesizer is able to repeat utterances precisely and without measurable variance. And (what a tape recorder cannot do) it can alter one or more parameters upon instruction, keeping all the rest the same.

But this requirement of complete and explicit specification can be a drawback too, especially for the study of prosodic features, because of the effort that must be spent in fussing over segmentals to get high-quality synthesis (this excludes all synthesis by rule to date). Control parameters for speech synthesizers have conventionally been treated as functions of time. A typical set of parameters would include the fundamental frequency, the frequencies (and, for parallel-resonator configurations, the amplitudes) of three or four variable formants, and the frequency and amplitude of hiss, with provision for hiss through formants. Of these, only fundamental frequency ($f_0$) is primarily a prosodic parameter, relating mainly to the accentuation and intonation of the utterance, with only comparatively small perturbations to simulate consonantal articulatory effects.

In the early machines the parameters were physically recorded as analog curves to be sensed and converted into appropriate control voltages by a special-purpose function generator. For intelligibility, time correspondence of the segmental parameter changes is critical. Thus it was natural to regard the control parameter curves as so many parallel channels with a common time-axis (the time scale often being 5 ips--i.e. that of the Sonagram supplying the formant frequency parameters). This approach carried over to synthesizers where the control parameters are quantized and stored on punched tape or in a digital computer. The control voltages produced by D-to-A conversion are immensely more reliable than those from the older hardware, but the parameters are still lock-stepped. Even though bandwidth compression efforts no longer form the chief motivation for synthetic speech research, the conceptual shackles they imposed remain with us.

A more sophisticated approach would differentiate between segmental parameters (including perhaps one labeled "$f_0$ perturbations") and prosodic parameters--e.g. intonation, rate, and several ones relating to voice quality--which can with impunity be rather loosely aligned with each other and with the segmental parameters. Incorporating "rate" as a parameter implies more than merely uniform stretching or compressing of the x-axis in time. It means that the parameter abscissae become abstract with respect to time--representing "articulatory progression through the utterance" or the like. The duration of any syllable (or any segment, for that matter) then is controlled by the instantaneous-rate-parameter value. In view of their tolerance of approximation by smoothed step functions of low bit-rate, the prosodic parameters could be abstracted still further in ways which would approximate prosodic synthesis by rule. The desideratum from the viewpoint of prosodemic research is to be able, once an utterance has been synthesized with good segmentals, to modify the prosodic stratum of synthesizer control parameters so as to produce various paradigms of accent shift and intonation switch.

## Artificially Manipulating Prosodic Features of Natural Speech

At the Center for Research on Language and Language Behavior we have developed means for introducing controlled changes into the prosodemic parameters of very natural-sounding speech without having to be concerned with synthesizing the segmentals at all. This technique, which we call "synthetic intonation," starts with human speech and alters the $f_0$ and durations of voiced segments by time-domain operations directly on the waveform. The formant frequencies and amplitudes of the voiced sounds are substantially unaffected.

Figure 1 shows the basic hardware configuration. A small (8k x 18 bit) laboratory computer samples and stores, displays, processes, and plays back speech samples. During the "record" phase the audio signal is input via an analog-to-digital converter which samples the waveform at a rate of nearly 8k samples per sec. with a resolution of 8 bits. Currently the computer can store about 7000 samples, or 7/8 sec. of telephone-quality speech. (Since the samples occupy less than half of each 18-bit word, this limit could be doubled by half-word storage were it not for our use of remaining bits for markers--see below.) Such 7000-element vectors form the basic observations of natural speech (typically spoken in a monotone) which after editing can be transformed by applying numerous different synthetic intonations. The playback phase in which the stored digital waveform is converted back to an analog signal can be performed without editing. The digital conversion rate is the same as the input sampling rate--nearly 8k samples per sec. The output signal is low-pass filtered at 4000 Hz to r ject high-frequency sampling noise. Unedited output of course provides a reference (manifesting all the effects of digitizing and filtering) against which to compare the output after prosodic modifications have been introduced in the "edit" phase.

Insert Figure 1 about here

For editing, the waveform is displayed on an x-y oscilloscope, 256 samples (or fewer, depending on scale magnification used) at a time. The scopeface "window" can be moved back or forth through the stored vector under potentiometer control. The middle of the window, shown by a dashed vertical line, pinpoints particular locations in the vector. At present the pitch marks are manually inserted (or erased) via the teletype keyboard at locations determined by the scope window centerline. Shortly we hope to automate the placement of these markers by the use of a hardware pitch extractor. Figure 2a shows the appearance of a waveform (for the vowel /a/) on the scope face before the insertion of pitch-period markers; 2b shows the same vowel with markers in place (the window being slightly offset).

Insert Figure 2 about here

The way in which the waveform is altered to realize a change in pitch is as follows: To increase pitch, a portion of each pitch period is "truncated." In other words, when the waveform is played back, a number of samples (proportional to the desired $f_o$ increase) are deleted or passed over in each pitch period. Conversely, to effect a decrease, each pitch period is "padded" by pausing briefly in the progression through the stored vector so that the same sample value is played back for two or more D-to-A cycles (the length of the interpolation being proportional to the desired $f_o$ decrease). These actions occur at the point in the waveform where a pitch-period marker is encountered.

Assume that K pitch-period markers, $m_j$, are positioned within the sample vector, where

$$0 \leq m_j \leq 7000, \qquad 1 \leq j \leq K,$$

and

$$m_j < m_{j+1} \qquad \text{for} \qquad 1 \leq j < K.$$

These markers then define K-1 pitch-periods, where the length of the $j^{th}$ period (in sample-size units) is:

$$l_j = m_{j+1} - m_j.$$

Once pitch-period markers are in place, the operator can specify a pitch contour for the utterance. This is done by partitioning the sample vector into arbitrary contiguous time intervals and specifying a pitch level for each boundary point. There are four pitch levels (numbered à la Trager and Smith, and defined in relation to the input $f_o$): 1 = lower, 2 = same, 3 = higher, and 4 = extra high. Potentiometers on the parameter control panel adjust the amount of lowering effected by pitch level 1 and of raising effected by pitch levels 3 and 4.

Consider the assignment of n $(n \leq 6)^2$ time-boundary markers, $t_i$, where

$$t_i = m_j \qquad \text{for some } m_j$$

such that

$$1 \leq i \leq n, \qquad 1 \leq j \leq K,$$

and

$$t_i < t_{i+1} \qquad \text{for} \qquad 1 \leq i < n.$$

For each $t_i$ there is an associated length-modifier,

$$p_i = P_h,\ 1 \leq h \leq 4,$$

where $P_1, \ldots, P_4$ correspond to four ascending relative pitch levels and are defined in terms of sample-size units such that

$$-p_{\ell im} \leq P_1 \leq 0,$$
$$P_2 = 0,$$
$$0 \leq P_3 \leq p_{\ell im},$$
$$P_3 \leq P_4 \leq p_{\ell im}$$

where $p_{\ell im}$ is some pragmatic limit on period-length alterations.

Although pitch values are assigned only at time boundaries, the program calculates a linear function between each pair of consecutive pitch assignments and alters the duration of individual pitch periods accordingly. In this way the period length modifications (and consequently, for monopitch speech input, the period lengths also) vary linearly along each vector interval bounded by consecutive time markers with associated pitch levels differing in value. Any periodicity perturbations in the data vector (including pre-existing intonations) of course add algebraically to the computed linear function, so that $f_0$ correlates of articulation and the slight aperiodicities of human phonation are preserved.

The modified length $\ell_i'$ of the pitch period beginning at $t_i$--i.e., the period over the interval $[m_j, m_{j+1}]$--for a given length-modifier assignment, $p_i$, is accordingly:

$$l_i' = l_i + p_i$$

where

$$l_i = m_{j+1} - m_j.$$

At time boundaries $t_i$ and $t_{i+1}$, the associated length-modifier values $p_i$ and $p_{i+1}$, respectively, define a linear function for calculating length-modification values $q_j$ for all intervening pitch periods over the interval $(t_i, t_{i+1})$:

$$q_j = am_j + b \qquad \text{for } t_i < m_j < t_{i+1},$$

where

$$a = \frac{p_i - p_{i+1}}{t_i - t_{i+1}}, \qquad b = \frac{t_i p_{i+1} - t_{i+1} p_i}{t_i - t_{i+1}}$$

The length, $l_j$, of the $j^{th}$ pitch period over this interval is computed:

$$l_j' = l_j + q_j.$$

Finally, the instantaneous fundamental frequency of the $j^{th}$ pitch period is related to its output length ($l_j'$, in sample-size units) by

$$f_{\circ j} = \frac{R}{l_j'}$$

where $f_{\circ j}$ is the fundamental frequency in Hertz corresponding to the $j^{th}$ pitch period; and

$R$ is the sampling/playback rate--approximately 8k samples/sec.

In inputting pitch-period markers during the "edit" phase, we try to place them (a) at closely corresponding points in consecutive periods, and (b) in regions of low variance--i.e. where the formant resonances are well damped. However, the speech quality turns out to be robust not only to variance in the location of markers but also to the distortion which our procrustean algorithm might theoretically be expected to introduce. Indeed it is in general not possible to discriminate pitch-altered syllables from ones played back at their original frequency (so subjected to the same digitization and filtering) unless extreme pitch changes or unnatural intonations have been imposed. Figure 3a shows output waveforms of a vowel /a/ corresponding to each of the four pitch levels. The input (pitch level 2) was at 100 Hz. Pitch level 1 was adjusted to be a major third below (80 Hz), and pitches 3 and 4, a minor third and minor sixth above, respectively (120 and 160 Hz). Thus the four tones form a major triad plus octave

Insert Figure 3 about here

(do, mi, sol, do') with the root in the bass at 80 Hz. Figure 3b shows the pitch-level-2 (unaltered) output obtained from four different inputs in which the same S produced /a/s naturally at these four pitches. Note that the two 100 Hz tokens of /a/--both produced naturally at that pitch--show waveshape differences of the same order of magnitude as those observable between the synthesized-versus-natural-pitch pairs, suggesting that the latter differences may be largely adventitious too. Synthetic pitch inflections are shown in Figure 4a (rising) and 4b (falling).

Insert Figure 4a and 4b about here

For illustrative convenience, these octave glides are shown spread over far fewer pitch periods than would occur in simulating natural speech. This was done by placing two time-boundary markers close together (10 pitch periods apart) and setting their associated pitch levels to (1,4) for rising and to (4,1) for falling.

Duration control. From the algorithm described above it should be evident that an output waveform contains all and only the pitch periods (sometimes truncated, sometimes "padded") which were present in the input. Consequently the durations of voiced portions are progressively shortened as the pitch is raised, and lengthened when the pitch is lowered. This dependence gives rise, as a by-product of introducing intonations synthetically, to rhythmic anomalies which are conspicuous with any but very strait (narrow) tessituras (i.e. small differences among the values $P_1, \ldots, P_4$) and become quite objectionable when the tessitura is spread toward the limits of $p_{lim}$. Accordingly, we are now developing a more sophisticated version of the program which will control durations independently of pitch. A simple approach would be to make the durations in the output match those in the input. This would involve straightforward compensation for the time removed from or inserted into pitch periods by the intonation algorithm. Where the $f_o$ is raised by truncating pitch periods, the duration would be normalized by reduplicating a pitch period whenever necessary. For example, if the $f_o$ were to be nearly doubled, then almost every pitch period would be played twice (in truncated form), as shown schematically in Figure 5b. Conversely, to keep the durations from growing longer with lowered $f_o$, pitch

---

Insert Figure 5 about here

---

periods would be deleted when required, as shown in Figure 5c. For certain purposes--e.g., for real-time manipulation of sidetone pitch (see below)--where correspondence of segmental durations in the output to those of the input is necessary or desirable, this approach would commend itself. But for synthetic intonation applications generally it would be preferable (just as for conventional speech synthesis; see above) to make "local rate" a separately controllable parameter, varying as a function of progression through the stored vector. The strategy would be basically the same as before, but--to take a rather extreme example--if a vowel spoken at 100 Hz were to be raised an octave (to 200 Hz) and simultaneously stretched to twice its original length, then each input pitch period (truncated to half its original length) would be played four times. It may be that multiple reduplication of natural human voice periods with their cycle-by-cycle fluctuations (cf. Kersta, Bricker, & David, 1960; Lieberman, 1961) will produce unacceptable pitch anomalies--this remains to be determined empirically.[3]

Treating "local rate" as an independent control parameter promises a potentially valuable spin-off in the form of a pitch-synchronous voice stretcher/compressor. The $f_o$ variations of voiced portions of the input speech sample would be duplicated at output but with the time axis altered (the time adjustment per period being inversely related to local $f_o$). We speak of this as a voice (not speech) stretcher/compressor because voiceless segments could not in general be expanded or shrunk in the same way as the voiced ones without anomalous consequences. (Fortunately, $p_{lim}$ protects voiceless intervals from being treated as gigantic pitch periods to be chopped or padded in the same ratio as the rest.)

The program is already operational as a voice stretcher/compressor for monotone speech in the sense that whenever a constant decrease or increase of pitch is assigned, there is an automatic duration effect. The results are quite promising.

## Projected Applications

Discourse synthesis. In addition to the application just discussed, there are several distinct ways in which "synthetic intonation" might function as a valuable tool for research and pedagogy relating to language and language behavior. The most central of these (that for which the program was initially intended) is "discourse synthesis." Speech samples of under a second have their uses, especially when interesting prosodic transformations can be wrought upon them. But for testing a model of English prosodemes, governed as they are by rules sensitive to contexts of greater than sentence length, one needs the ability to synthesize the prosodic parameters of relatively long discouses or texts--e.g. the 45 sec. "north Wind and the Sun" passage done on the Parametric Artificial Talker "PAT" at Edinburgh (Uldall & Anthony, 1962).

The CRLLB computer has a digital magnetic tape unit as one of its peripherals. But quite aside from its intermittent recusancy, it is unsatisfactory as a means of extending our synthetic intonation storage capacity, because whenever it writes out of core (or reads in) it slows the central processor by "cycle stealing." Since our hoary computer operates with an 8 μsec. cycle time, it must go at full tilt to digitize and store 8-bit samples from a single analog channel at 8k samples per sec. Simultaneously writing onto tape clobbers this sampling rate, and as Cooper (1963) notes, "the quality of pulse-code-modulation speech deteriorates rapidly with decreasing bit rate [p. 333]." Then when the tape record is finished, the cycle stealing stops, and the sampling rate goes back up again. We see no imminently implementable way around this except to input whole discourses from audio tape in 7/8 sec. chunks and, after processing each, dump them back onto another audio tape to be hand-spliced together afterward. Needless to say, this enterprise is not high on our priority list.

Prototype ultra-high-performance word-reading machine for the blind. While the 7/8 sec. restriction looks short with respect to texts, it is generous with respect to a vast majority of English words. Cooper (1963) discusses two approaches to the problem of speech output from reading machines for the blind: "One generates a spoken output by rearranging voice recordings of individual words into sequences found in the printed text; the other generates synthetic speech on the basis of letter-by-letter information from the text [p. 326]." The advantages of the word-reading machine with human voice recordings--more intelligible, lifelike speech, etc.--are offset not only by the formidable memory requirements (over 100 million bits for a 10,000-word vocabulary: Cooper, p. 333) but also by the barrier imposed against providing appropriate accents and intonations or giving blind users any choice of average rate. The letter machine using speech synthesis (or a word machine scoring synthesizer parameters instead of the audio waveform for each word, to save memory) could provide these variations, but at a cost of marginal speech quality and intelligibility.

Synthetic intonation appears to offer the best of both worlds (without ameliorating the memory requirements--which, however, no longer seem so staggering): the words are intelligible and lifelike, accents and intonations can easily be superimposed, and the average rate can be adjusted.[4] We hope to build up a modest vocabulary of monopitched words on digital tape and experiment with giving them the prosodic modifications appropriate to various positions in sentences of divers types. The machine would of course not operate in real time, and the problem of physically concatenating the audio records so produced will remain. A start-stop analog recorder (cf. Cooper, 1963) would be the method of choice for overcoming this.

Psycholinguistic test-item production. There are many occasions in the study of language behavior for using short verbal stimuli that sound like natural speech but are controlled or normalized as to crucial prosodic attributes such as length, SPL, and $f_o$. For example several dichotic listneing experiments have used dual track recordings of a succession of monosyllable pairs (the words in each pair being played simultaneously to opposite ears of the Ss) to determine, e.g., which ear will most often dominate the identification response. But the paired stimuli have typically exhibited substantial variance in timing, duration, and pitch--whose effects cannot be discounted, especially where they are not known to be random and symmetrical. The synthetic intonation program, with duration control added, will permit the pitch, length, and onset time of such stimuli to be controlled with the same precision as would routinely be done for SPL. Slight software modifications will split the storage vector into two parts each accomodating a word of less than one-half second, with corresponding portions of each word displayed in the scopeface window simultaneously. The two stimuli then could even be normalized on a period-by-period basis (provided they did not contrast voiced/voiceless or plosive/implosive consonants where $f_o$ is a cue to the manner of articulation or the airstream initiation). The pair of processed words would finally be recorded simultaneously onto two tracks of an analog tape.

Many of the listening experiments whose results have been viewed as supporting a "motor theory" of speech perception have used synthetic speech stimuli, which could be produced in a precisely controlled way and varied along orthogonal parameters. A frequent criticism, however, has been directed at the quality of the synthesized speech. For example, when Liberman, Harris, Kinney, and Lane (1961) studied categorical perception and labeling of seven Pattern Playback stimuli ranging perceptually from /do/ to /to/, typical Ss labeled even the two most extreme /do/-like stimuli as /to/ 5% of the time. One can infer from this (and confirm by listening to the tapes) that the approximation to natural speech was rather poor.

One of the few lines of inquiry viewed by the experimenters as favoring the motor theory of speech perception which has used natural rather than synthetic speech stimuli concerns the loudness function for heard speech (notably Peterson & Lehiste, 1959; Ladefoged & McKinney, 1963; for discussion see Lane, 1965). The hypothesis is that Ss relate their judgements of the loudness of others' speech not to its SPL directly but to the effort they themselves would have to make to produce such a sound. In these experiments pitch and duration, as well as spectral cues of vocal effort, were uncontrolled. Use of stimuli full of adventitious cues in addition to, and highly correlated with, the experimental variable vitiates the results of these studies so far as "proving" that listeners perceive speech by redintegrating to speech production.

Synthetic intonation offers the means of producing stimuli with which to replicate crucial parts of the earlier studies using a polydimensional experimental design (cf. Lane, 1962) in which the contributions of $f_o$, duration, SPL, vowel type, and glottal spectrum are separately evaluated (the first three parameters can be altered synthetically; the other two can be separated if the talker produces all vowel types at all levels of effort). We hope that such a controlled experiment will lay at least one ghost of a factual basis for the "motor theory."

Sidetone-pitch manipulation. In the applications discussed so far the recording and playback phases can be carried out separately and the problem is typically one of storage capacity. Another whole class of applications would use the computer and the synthetic intonation algorithms to modify prosodic parameters of live speech in real-time--i.e. with short-term buffering only, such that the delay introduced is small with respect to that which produces the characteristic effects of delayed auditory feedback (Yates, 1963). Several experiments suggest themselves, given the facility to intervene in Ss' auditory feedback loops in ways more sophisticated than can be done using gross masking noises, delays, and gain changes. For example, depressed speech is often characterized by flat, monotonous intonation contours (straitened tessitura).[5] If a depressed S's intonational variations (from his mean $f_o$) were multiplied by some factor, and his speech (processed accordingly) fed into his earphones at high gain, what effect would this have on his affect? It is well known (since e.g. James & Lange) that the causal link between affect and vocal index is a two way street, so quite possibly S would become less depressed from hearing his own voice with widened tessitura. A different prediction would emerge from the "servo-theory" of speech production: on the assumption that S unconsciously narrows his tessitura to express his affect, he could be expected to react to artificially widened feedback by a compensatory narrowing still further of his production. There is also a possibility that such a closed-loop system, especially when operated at greater-than-unity gain (in the frequency domain--i.e. multiplying deviations from mean $f_o$ by more than 1.0) will be inherently unstable, resulting in oscillatory behavior. These are empirical questions which await experimentation.

Unfortunately, our present computer is too slow to sample the speech wave and pitch meter pulses, process the signal, and play it back all in real-time. However, we are developing an interim program which will take audio tape input played at 1/4 its recorded rate, process it, and output onto another audio tape (to be subsequently played back at a correspondingly stepped-up rate). Such a non-real-time simulation will not manifest Ss' responses, of course, but will permit the program to be debugged and evaluated for use on faster hardware. In particular it will show whether, in the absence of feedback effects upon the input (the speaker's production), depressed speech for instance, of which we have good tape recorded examples, can be made to sound convincingly less depressed by the simple $f_o$ multiplication algorithm suggested above, or whether perhaps a much more complex function will be required. This question can best be answered in open-loop mode.

Mechanical pedagogy. Conventional language laboratory exercises aiming to establish target-language "stress" patterns seldom acheive this goal: "students reliably 'underestimate' the required loudness ratio [Lane, 1962, p. 16]." One feature obviously missing from the tape recorded aural-oral paradigm that is found in human teachers' behavior when eliciting successive approximations in a student's imitative vocal responses, is exaggeration of the lacking attribute. The SAID system presently implemented at CRLLB (Speech Auto-Instructional Device-Buiten & Lane, 1965) is able to track several prosodic parameters of a student's imitative response and signal the degree of acceptability achieved. We think that in principle (i.e. not with our present computing hardware) the pedagogical effectiveness of this system could be improved enormously by giving the feedback in the form of a repetition of the stored target speech pattern (or of the student's response), modififed by synthetic intonation algorithms so as to exaggerate whatever pitch, amplitude, or durational contrast the student failed to produce adequately.

Speech distortion. So far, applications involving alteration of natural-speech duration and/or fundamental frequency have been discussed. But the technique of separating $f_o$ from formant frequencies by pitch synchronous, time-domain operations can be applied more generally. By setting the playback rate unequal to the sampling rate, the formant frequencies of voiced portions could be altered while, with compensatory truncation or padding of the pitch periods, the $f_o$ were kept constant. Raising the formant frequencies of speech (without raising its rate and $f_o$ as in speeding up a tape recording) would simulate the well known effects of "helium speech" though not the high pressure effects also found in divers' speech (cf. Brubaker & Wurst, 1968; Fant & Lindqvist, 1968; Gerstman, Gamertsfelder, & Goldberger, 1966). The converse transformation (lowering the formants) would appear to offer a simple time-domain approach for automatically enhancing the intelligibility of divers' speech.

## Summary

Synthetic intonation refers to means for imposing controlled changes on certain prosodic parameters of natural human speech. In particular, fundamental frequency and local rate can be controlled by pitch-synchronous, time-domain

operations on digitized speech waveforms. To raise the pitch, each period is truncated by skipping one or more consecutive sample values in playback. To lower the pitch, each period is extended by playing back one sample value for two or more time units.

For basic research into the prosodemic structure of languages, and for applications such as an ultra-high-performance word-reading machine for the blind, producing controlled stimuli for psycholinguistic experiments, and sophisticated teaching-machine pedagogy, synthetic intonation offers advantages over conventional speech synthesis by art (convenience) and by rule (speech quality and intelligibility). The dawning era of machine-man communication holds promise of a substantial role for synthetic intonation.

Footnotes

[1]The research reported herein was performed in part pursuant to Contract OEC-3-6-061784-0508 with the U. S. Department of Health, Education, and Welfare, Office of Education, under the provisions of P. L. 83-531, Cooperative Research, and the provisions of Title VI, P. L. 85-864, as amended. This research report is one of several which have been submitted to the Office of Education as Studies in Language and Language Behavior, Progress Report VIII, February 1, 1969.

[2]The restriction of n to six or fewer time boundaries is entirely system dependent and applies only to our preliminary implementation where a limited array of toggle switches stores the associated pitch levels. We hope ultimately to derive time markers automatically from acoustically manifested syllable boundaries (cf. pitch markers from $f_0$ extraction) and to be able, by specifying minimal prosodemic units, to impose appropriate trapezoidal (or perhaps trapezial) $f_0$ functions (i.e., integrated step-functions).

[3]Although our algorithm expressly preserves such periodicity variance, or jitter, it could of course equally well be revised to normalize adjacent pitch-period lengths at the expense of making the output speech sound more "machine-like" (cf. Kersta, Bricker, & David). Alternatively the jitter could be preserved in a more natural way during voice stretching by reduplicating not single periods but spans of, say, four.

[4]Natural speech could be employed also, with the same relative benefits, for segment concatenation synthesis. The stored segments would have fixed, monopitch $f_o$; and appropriate prosodic patterns would be imposed on the output.

[5]The term "lack of affect" sometimes used for the speech of depressed persons seems to us both misleading and erroneous. It is misleading because it confounds affect proper with vocal indices (cf. Abercrombie, 1967) of affective states; it is erroneous because it implies that either depressed speech is indistinguishable from neutral speech of normals, or else normals perpetually exhibit affective indices.

## Figure Captions

Fig. 1 Block diagram of hardware configuration for synthetic intonation.

Fig. 2 (a) View of scopeface showing digitized waveform of vowel /a/. Vertical line at center screen locates markers.

(b) Same waveform with pitch period markers in place; display window has been offset four samples left.

Fig. 3. Comparison of synthetic and natural pitch. Oscillograms of output waveforms of /a/ at four fundamental frequencies: (a) imput at 100 Hz, other output frequencies produced synthetically; (b) input at 80, 100, 120, and 160 Hz; output unaltered.

Fig. 4. Rising and falling voice frequency inflections, artifically introduced. Octave changes are shown compressed for illustration into spans of ten pitch periods.

Fig. 5. Schematic representation of input (a) and output (b, c) waveform envelopes, showing: (b) reduplication of truncated periods, and (c) deletion of excess periods, to preserve durations independent of pitch.

References

Abercrombie, D. Elements of general phonetics. Chicago: Aldine, 1967.

Brubaker, R. S., & Wurst, J. W. Spectrographic analysis of divers' speech during decompression. Journal of the Acoustical Society of America, 1968, 43, 798-802.

Buiten, R., & Lane, H. L. A self-instructional device for conditioning accurate prosody. IRAL, 1965, 3, 205-219.

Cooper, F. S. Toward a high performance reading machine for the blind. In E. M. Bennett, J. Degan, & J. Spiegel (Eds.), Human factors in technology. New York: McGraw-Hill, 1963, 326-334.

Fairbanks, G. A theory of the speech mechanism as a servosystem. Journal of Speech and Hearing Disorders, 1954, 19, 133-139.

Fant, G., & Lindqvist, J. Pressure and gas mixture effects on diver's speech. Quarterly Progress and Status Report, No. 1/1968. Stockholm: KTH Speech Transmission Laboratory, 1968, 7-17.

Gerstman, L. J., Gamertsfelder, G. R., & Goldberger, A. Breathing mixture and depth as separate effects on helium speech. Journal of the Acoustical Society of America, 1966, 40, 1283. (Abstract)

Kersta, L. G., Bricker, P. D., & David, E. E., Jr. Human or machine?--a study of voice naturalness. Journal of the Acoustical Society of America, 1960, 32, 1502. (Abstract)

Ladefoged, P., & McKinney, N. P. Loudness, sound pressure, and subglottal pressure in speech. Journal of the Acoustical Society of America, 1963, 35, 454-460.

Lane, H. Psychophysical parameters of vowel perception. Psychological Monographs: General and Applied, 1962, 76(44) Whole No. 563), 1-25.

Lane, H. The motor theory of speech perception: A critical review. Psychological Review, 1965, 72, 275-309.

Lehiste, I., & Peterson, G. E. Vowel amplitude and phonemic stress in American English. Journal of the Acoustical Society of America, 1959, 31, 428-435.

Liberman, A. M., Harris, K. S., Kinney, J., & Lane, H. L. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology, 1961, 61, 379-388.

Liberman, P. Perturbations in vocal pitch. Journal of the Acoustical Society of America, 1961, 33, 597-603.

Uldall, E., & Anthony, J. K. The synthesis of a long piece of connected speech on PAT. Edinburgh University Phonetics Laboratory report, n. d. (c. 1962).

Yates, A. J. Delayed auditory feedback. Psychological Bulletin, 1963, 60, 213-232.

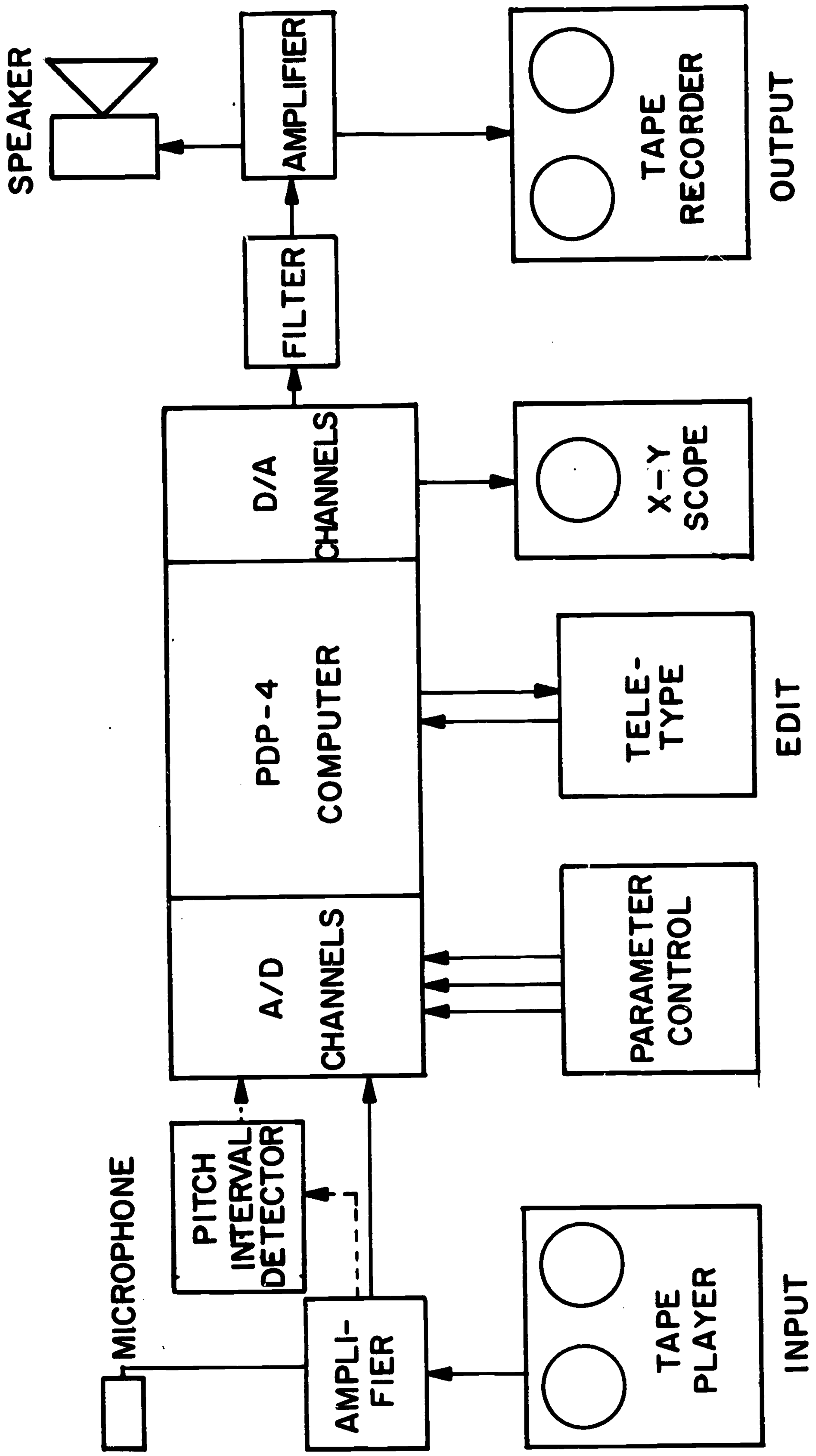
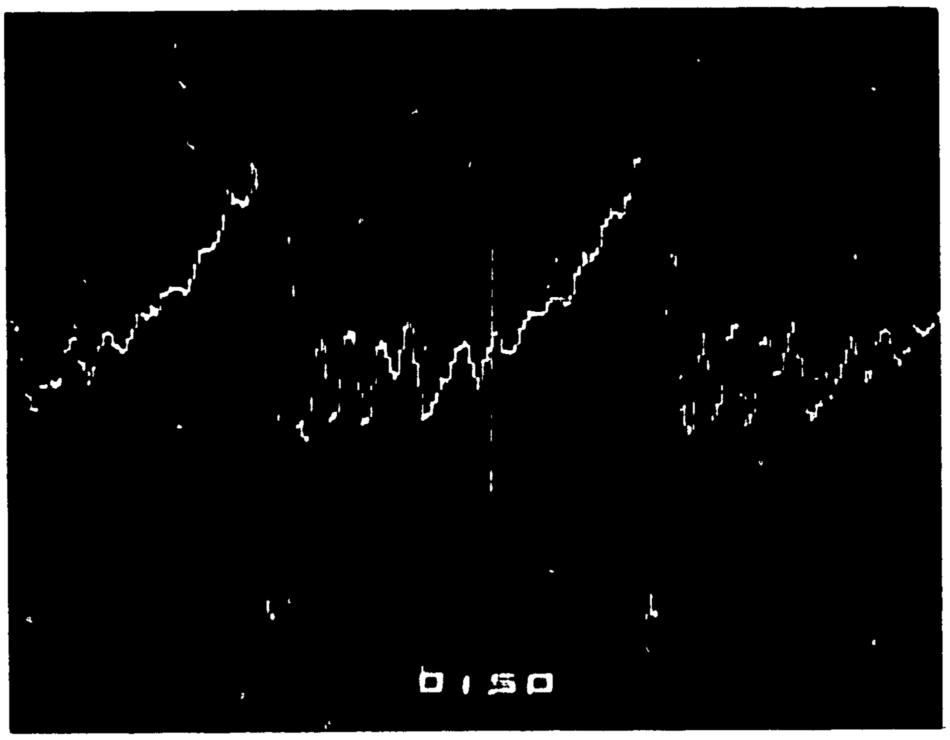
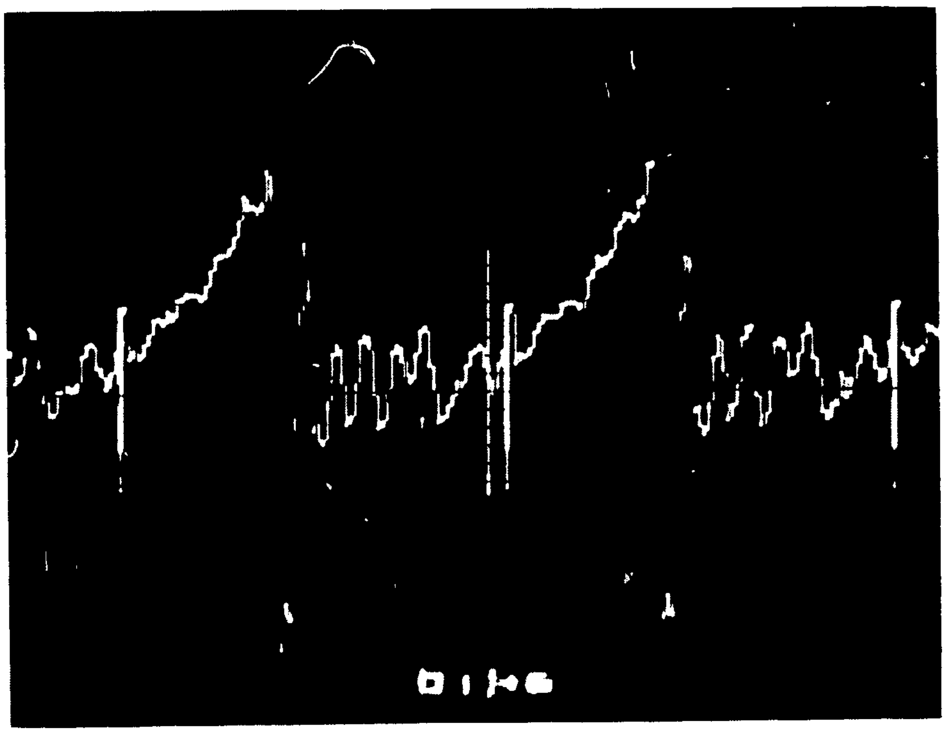

SPEAKER
AMPLIFIER
TAPE RECORDER
OUTPUT
FILTER
D/A CHANNELS
X–Y SCOPE
PDP-4 COMPUTER
TELE-TYPE
EDIT
A/D CHANNELS
PARAMETER CONTROL
MICROPHONE
PITCH INTERVAL DETECTOR
AMPLI-FIER
TAPE PLAYER
INPUT

Figure 1




(a)



(b)

Figure 2

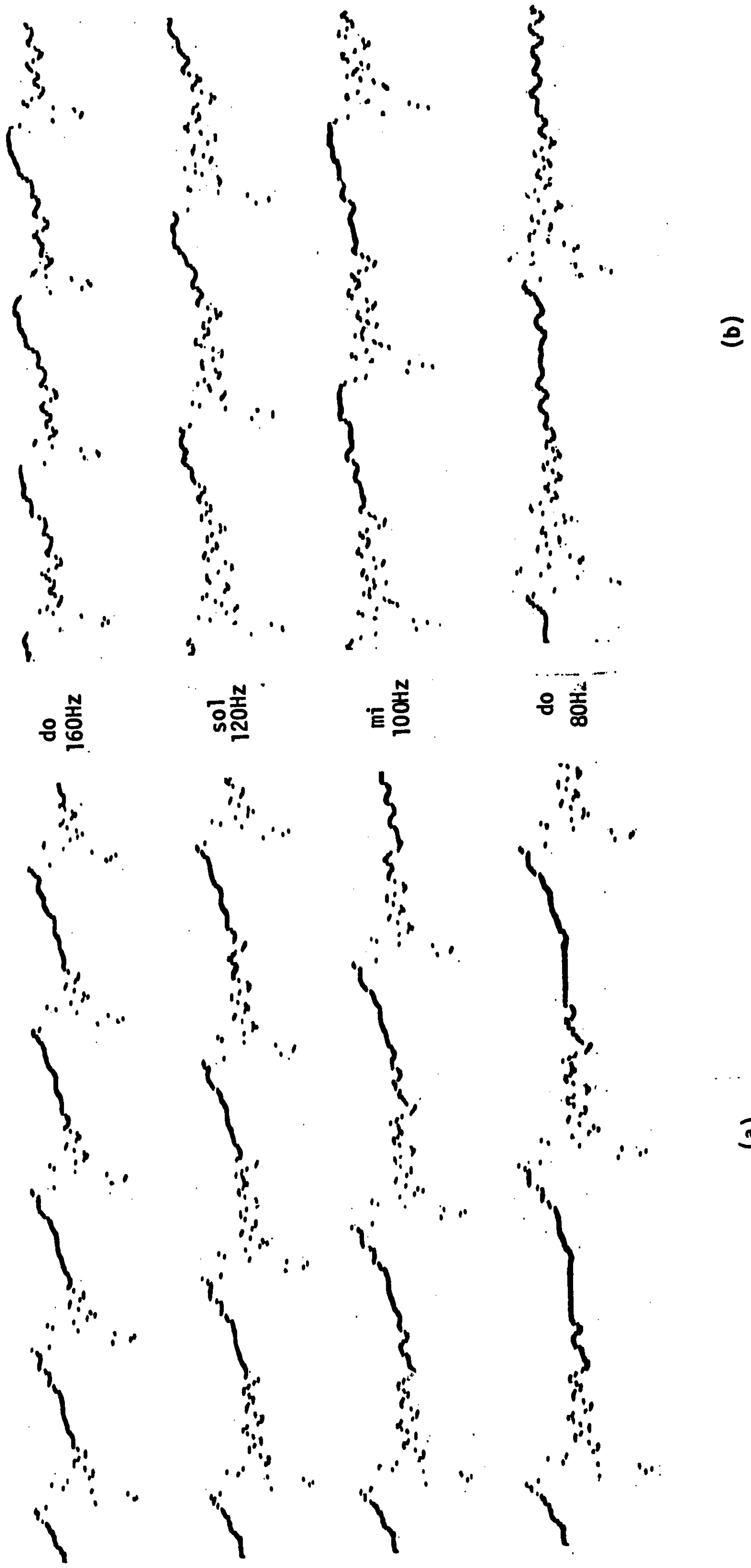
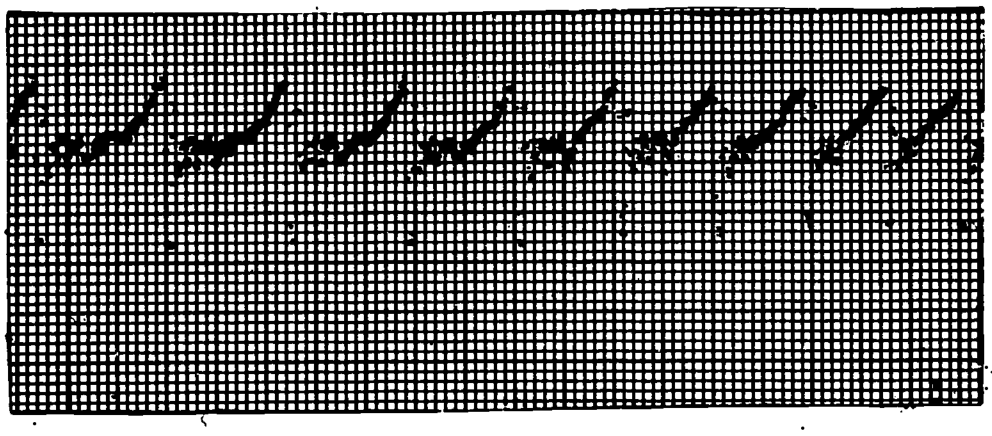
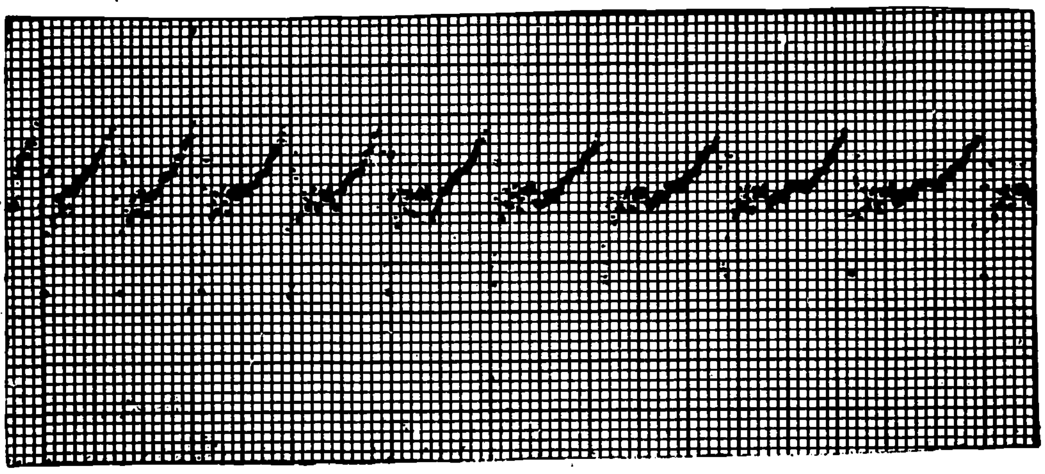



Figure 3



(a)



(b)

Figure 4

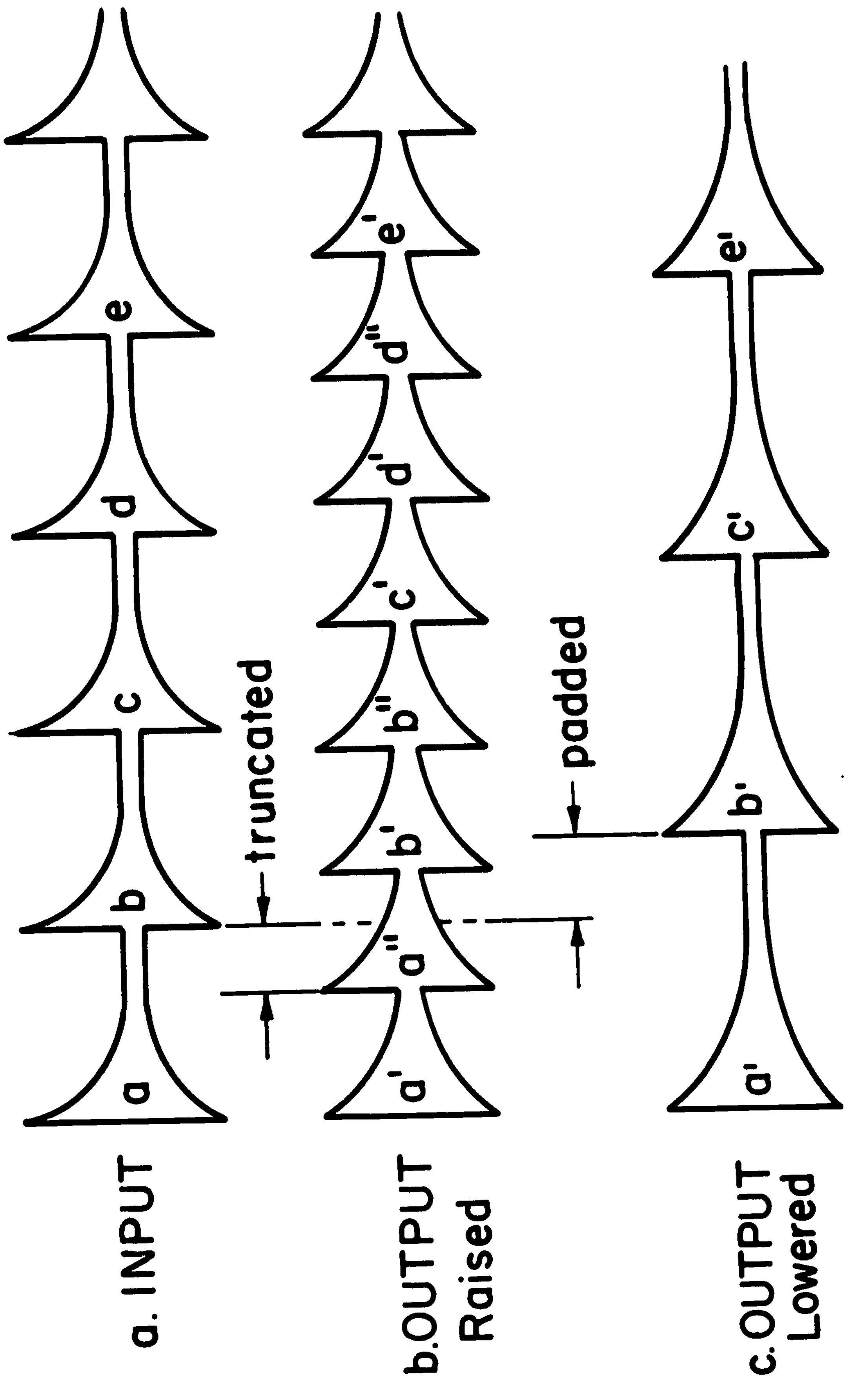

a. INPUT
b.OUTPUT
Raised
c.OUTPUT
Lowered
truncated
padded
a
b
c
d
e
a'
a''
b'
b''
c'
d'
d''
e'
a'
b'
c'
e'

Figure 5