

AU-A138 660

VOCODER ANALYSIS BASED ON PROPERTIES OF THE HUMAN
AUDITORY SYSTEM(U) MASSACHUSETTS INST OF TECH LEXINGTON
LINCOLN LAB B GOLD ET AL. 22 DEC 83 TR-670

1/1

UNCLASSIFIED

ESD-TR-83-226 F19628-80-C-0002

F/G 5/10

NL

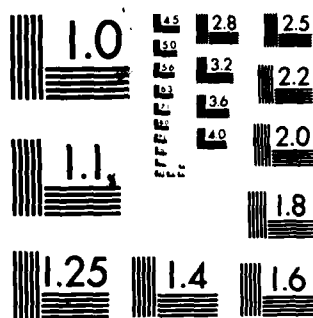
END

DATE

FILED

4-84

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

VOCODER ANALYSIS BASED ON PROPERTIES
OF THE HUMAN AUDITORY SYSTEM

B. GOLD
J. TIERNEY
Group 24

TECHNICAL REPORT 670

22 DECEMBER 1983

Approved for public release; distribution unlimited.

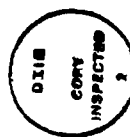
DTIC
ELECTE
S MAR 6 1984 D
B

LEXINGTON

MASSACHUSETTS

ABSTRACT

When a person listens to speech corrupted by noise or other adverse environmental factors, speech intelligibility may be impaired slightly or not at all. The same corrupted speech, after being vocoded, often causes drastic intelligibility loss. This is due to the fact that the human peripheral auditory system is a superior signal processor to that of the vocoder. This report is based on the premise that a vocoder analyzer that better resembles the peripheral auditory system would function in a superior manner to present-day vocoders. Topics include reviews of speech enhancement techniques, perceptual analysis of diagnostic rhyme test data, a brief description of the peripheral auditory system and an outline of proposed psychophysical tests. The final section is devoted to a discussion of some preliminary work on computer simulation of an auditory model.



111

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

CONTENTS

Abstract	111
I. INTRODUCTION	1
II. SPEECH ENHANCEMENT TECHNIQUES	2
III. PERCEPTUAL ANALYSIS OF VOCODERS BASED ON DIAGNOSTIC RHYME TEST DATA	4
IV. PERCEPTUAL ANALYSIS OF VOCODERS BASED ON PITCH EXPERIMENTS	24
V. THE PERIPHERAL AUDITORY SYSTEM	35
VI. MODELLING THE AUDITORY SYSTEM	56
REFERENCES	66
APPENDIX A	69
APPENDIX B	77
APPENDIX C	83

ILLUSTRATIONS

1. Intelligibility vs. cut-off frequency for "k" (from Miller-Nicely data).	10
2. Intelligibility vs. cut-off frequency (from Miller-Nicely data), all 16 consonants.	11
3. Total number of correct responses vs. s/n for speech bandwidth 200-6500 Hz (from Miller-Nicely).	12
4. Aggravation factor for four systems.	21
5. DRT scores vs. systems in F-15 noise.	22
6. Cartoon-like representation of the peripheral auditory system.	25
7. Block diagram of energy transfers in the ear.	26
8. Some views of the auditory system. (a) Sectional and perspective views of the human hearing mechanism. (b) Sectional view of the cochlea. (c) Schematic view of the human hearing mechanism showing the outer ear, the middle ear, the cochlea, and the nerve fibers leading to the brain.	27
9. Block diagram of peripheral auditory system. From Weiner (1949).	38
10. The major structural features of the uncoiled cochlea. Note that the basilar membrane is narrow near the round window and wider near the helicotrema, a taper opposite the cross-section area of the cochlea.	40
11. Structural and anatomical features of the cochlea. (a) The cochlea in relation to the middle ear and auditory nerve. (b) Cross section of the cochlea. (c) The scale media (from Green "An Introduction to Hearing").	41
12. Computer model of basilar-membrane displacement. Response to a single rarefaction impulse of sound pressure at the eardrum.	42
13. Analytical model for basilar-membrane displacement. (After Flanagan, 1962).	44

ILLUSTRATIONS (continued)

14. Camera lucida drawing of a cross section of the cochlea partition in the second turn of a guinea pig cochlea. The attachment shown here of the tectorial membrane to the inner supporting cell, and to Hensen's cells, is based on microdissection of fresh, unfixed specimens. From Davis (1961). 46
15. Relationship between the tectorial membrane and cilia of outer hair cells. At rest (lower illustration) the cilia stand perpendicular to the cuticular surface of the cell. When pressure waves move the basilar membrane, a "shearing" force acts to alter the angle of the cilia with respect to the cuticular surface. Note that the cilia of the inner hair cells are shown to bend, not from tectorial membrane attachment but from fluid motion. From P. Dallos and A. Ryan, : Physiology of the inner ear. In J. L. Northern (ed).: Hearing Disorders, 1976, p. 95, (Little Brown and Co.). 47
16. Interval histograms for tone bursts at different frequencies. 50
17. Diagram of the auditory pathways linking the brain with the ear. 52
18. Diagrammatic representation of membrane potential, threshold potential, and spike activity of the model neuron. From Weiss. 59
 R_M =maximum threshold potential
 R_R =resting threshold potential
 T_R =time constant of the exponential decay
of the threshold from its maximum to its
resting value
19. (a) Block diagram of the cascade/parallel filterbank.
(b) Pole-zero plots and transfer functions of filters used in the filterbanks.
(c) Block diagram of one channel of the detection and compression models. 61
20. Neuron-like elements in pitch measurement model.
(a) Nomenclature for simple "neuron" N.
(b) Operation of neuron N.
(c) Level B of neural net. 63

21. Outputs of several filters and neurons for male utterance of length 150 msec. (See text for full explanation.)	65
A-1. Number of correct plosive identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).	71
A-2. Number of correct fricative identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).	72
A-3. Number of correct nasal identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).	73

I. INTRODUCTION

When speech is degraded before entering the listener's ear, the resultant signal is generally less intelligible. However, loss of intelligibility is made worse when the degraded speech is first processed by any of a great many speech communications systems. This fact is of great practical concern, for example, in a military aircraft environment, where intelligible communications can become a life and death matter. Also, under certain adverse conditions (such as a jamming environment), low bit-rate speech is mandatory. Thus, the fact that present-day operational vocoder systems greatly reduce intelligibility of speech further aggravates an already difficult situation. This report summarizes recent work at Lincoln Laboratory dedicated to better understanding this problem with the hope of alleviating it. Our approach is strongly oriented towards exploiting the properties of human auditory perception and physiology. Such an approach can prove to be beneficial for the following reason:

The prevalent model of the peripheral human auditory system resembles a bank of many highly overlapped bandpass filters. The particular shape of these filters, the associated non-linearities and the relation of these anatomical and physiological models to psychophysical experiments have been the subject of intensive study. Existing VLSI technology makes the engineering implementation of crude auditory models quite feasible. Since there is evidence that processing by the human auditory system is less inimical to intelligibility than is a vocoder, these models may strongly suggest methods of designing more robust future vocoders.

The report will be divided as follows: in Section II we review the existing knowledge on speech enhancement techniques. To a great extent, these methods deal with degradations caused by additive acoustic noise and focus on noise reduction techniques. Unfortunately, all such attempts also remove important information-bearing elements from the speech wave and generally result in no improvement of intelligibility. Next, in Section III we present evidence that vocoding generally aggravates intelligibility loss, and speculate as to some of the probable causes. In that section, our main evidence will be the body of knowledge contained in recently obtained DRT (Diagnostic Rhyme Test) results. In Section IV we review a few of the major results of the psychophysics of human pitch perception and present some of our recent work directed towards extending these results to pitch perception of speech. In Section V we review and synopsise some of the important knowledge gained about the functioning of the human peripheral auditory system. Finally, in Section VI we present preliminary results on a computer simulation program of a model of the peripheral auditory system.

II. SPEECH ENHANCEMENT TECHNIQUES

The book Speech Enhancement edited by Jae S. Lim [1] contains much useful information on reported efforts to reduce acoustic noise or mitigate its effects. In his overview to part one of the book, Lim summarizes progress to date:

In summary, successful results are available only in rather restricted applications. With narrowband background noise, for example, simple linear filtering can significantly improve speech intelligibility. For wideband random background

noise, speech quality can be improved by various algorithms discussed in this section. In the context of bandwidth compression of noisy speech, speech quality can be improved in the presence of wideband random noise, and some studies suggest that small improvements in intelligibility may also be possible. Major unsolved problems remain, however. For example, no algorithm has been shown to improve speech intelligibility when speech is degraded by wideband random noise and there is only one microphone input. In the case of interference from competing speakers, improvement has not been demonstrated in either intelligibility or quality. Even with multiple microphone inputs, significant intelligibility improvement has been demonstrated only in restricted environments.

The message contained in this paragraph is that noise cannot easily be stripped away from speech without seriously compromising the speech intelligibility. Much of the work in part one of Lim's book is related to classical work on detection of simple signals (such as sine waves) in noise or relate, at least philosophically, to Wiener's theoretical ideas. But speech is neither a simple signal nor a statistically stationary signal.

At this point we should inquire as to how capable are people of extracting speech in noise. Quantitative information on this subject is sparse. Nevertheless, most of us will agree with the "conventional wisdom" that people do well in this task. Cherry [2] has summarized some ideas on this subject via discussion of the "cocktail party effect" as follows:

(1) At a loud cocktail party, the listener can watch the speaker's mouth and gain some speech cues by purely visual means.

(2) Binaural hearing permits the listener to direct his attention at the sounds emanating from the speaker's mouth.

(3) Syntax and semantics of the conversation are obviously of great help.

(4) Tracking of the speaker's pitch and timbre permit some "blocking out" of the noise.

Part I of Lim's compilation deals with enhancement of speech degraded by additive noise. The methods used can be categorized as:

- (a) methods for trying to improve the signal to noise ratio;
- (b) methods for trying to take advantage of the periodicity of speech;
- (c) methods for trying to take advantage of our knowledge of the speech production model;
- (d) methods using more than one microphone input.

Thus far, none of the above methods have demonstrated any improvement in intelligibility.

Part II of Speech Enhancement deals with processing methods that are applied to the speech prior to its degradation by additive noise. Thus, these papers have no relevance to the problem of an acoustic noise environment.

III. PERCEPTUAL ANALYSIS OF VOCODERS BASED ON DIAGNOSTIC RHYME TEST DATA

Since 1975, many speech processing systems have been subjected to Diagnostic Rhyme Tests (DRTs) and much interesting information is available from these tests. For example, the DRT averaged over three speakers for speech with no acoustic noise background is 98.4 and this score drops to 92.6 when appreciable simulated aircraft noise is acoustically added. This is a loss of 5.8 points, not insignificant. On the other hand, the DRT for

Lincoln Laboratory LPC-10 is 88.1% and drops to 75.2% with the same noise added, a loss of 12.9 points. Thus, the LPC-10 processing algorithm "aggravates" the DRT loss by $12.9 - 5.8 = 7.1$ points.

What causes a vocoder to aggravate intelligibility loss for noisy input speech? One may speculate as follows:

(1) Much of the information in the speech wave comes from the consonants, which are typically low level sounds so they can be more easily masked by the noise.

(2) A typical spectral cross section of a vowel, glide or nasal has peaks and valleys. The valleys may be masked even though the peaks have good speech to noise ratio.

(3) During a vowel sound, the speech energy is concentrated in narrow bands about the harmonics of the fundamental, while noise is usually spread over the entire spectrum. During voiced sounds, a vocoder tends to transform this noise into modulation on the speech harmonics.

DRT scores are a convenient way of comparing systems for two reasons. First, they are of widespread use in rating government speech systems. Second, the data obtained can be used as a diagnostic tool.

The Diagnostic Rhyme Test (DRT) has, for the past few years, been perhaps the most prevalent method of evaluating vocoder systems for DoD use [3,4]. In this section, we make further use of the DRT to study several important questions related to vocoder performance:

(a) The relative contributions of the vocoder pitch and spectral tracks to intelligibility loss.

(b) The "aggravation factor" of various systems, defined as the additional loss in DRT points caused by a vocoder in an acoustic noise background.

(c) The improvement in intelligibility due to various hybrid configurations wherein a portion of the unvoiced speech is added to the vocoded speech.

Confusion Matrices

DRT results are useful not only for obtaining relative scores for different cases but also because of the details of diagnosis. We have chosen to present these details via a confusion matrix; this matrix yields valuable information on the type of errors caused by a given system.

The confusion matrix was proposed by Miller and Nicely [5]. Their matrix consisted of a 16 x 16 array of consonants, as shown for example, in Table I (their Table II), reproduced below. By scanning any row (for example, the f row), one can see how many times f was mislabelled by the listeners and how it was mislabelled. For example, "f" was incorrectly identified as "p" 31 times, while it was correctly identified 85 times.

The Miller-Nicely data is based on utterances of each consonant followed by the vowel "ah" (as in father). DRT confusion matrices are based on the initial consonant-vowel combination of each DRT word. Thus, for example, if the word "daunt" is mislabelled as "taunt," this would correspond to a d+t confusion.

The Miller-Nicely processing of the utterances consisted of filtering and additive noise. Thus, in Table I, the signal to noise ratio was set for -12 db and the filter used covered the range 200-6500 Hz.

TABLE I

EXAMPLE OF MILLER-NICELY CONFUSION MATRIX

	p	t	k	f	θ	s	∫	b	d	g	v	δ	z	3	m	n
p	51	53	65	22	19	6	11	2		2	3	3	1	5	8	5
t	64	57	74	20	24	22	14	2	3	1	1	2	1	1	5	1
k	50	42	62	22	18	16	11	4	1	1	1	2			4	2
f	31	22	28	85	34	25	11	3	5		8	8	3		3	
θ	26	22	25	63	45	27	12	6	9	3	11	9	3	2	7	2
s	16	15	16	33	24	53	48	3	5	6	3	1	6	2	1	1
∫	23	32	20	14	27	25	115	1	4	5	3		6	3	4	2
b	4	2	2	18	7	7	1	60	18	18	44	25	14	6	20	10
d	3		1	4	7	4	11	18	48	35	16	24	26	14	9	12
g	3	1	1	1	4	5	7	20	38	29	16	29	29	38	10	9
v		1	1	12	5	4	5	37	20	23	71	16	14	4	14	9
δ		1	4	17	2	3	2	53	31	25	50	33	23	5	13	6
z	6	1	2	2	6	14	8	23	29	27	24	19	40	26	3	6
3	3	2	2	1		6	7	7	30	23	9	7	39	77	5	14
m		1			1	1		11	3	6	8	11		1	109	60
n	1			1		1		2	2	6	7	1	1	9	84	145

Table II lists the 17 different conditions used by Miller and Nicely. Examination of the resulting 17 confusion matrices permits us to evaluate how the 16 consonants are affected by various filtering conditions and for various noise conditions. It is instructive, for example to examine the intelligibility of each consonant in response to low pass and high pass filtering. In Appendix A, these effects are plotted separately for each consonant. For purposes of illustration, the graph for the consonant "k" is reproduced here as Fig. 1. The solid curve is derived from items 7 through 12 in Table II; the dashed curve is derived from items 13 through 17 of Table II. The ordinate represents the correct number of identifications while the abscissa represents the low cutoff frequency (for the solid curve) or the high cutoff frequency (for the dashed curve). Any horizontal cut shows a comparison of high pass versus low pass cutoff for equal intelligibility. Thus, in Fig. 1 a low pass filter from 200-1200 Hz gives the same intelligibility as does a high pass filter from 1700-5000 Hz.

Figure 2 is a composite graph of all 16 consonants. It is seen that the curves intersect at about 1500 Hz; this tells us that, for the Miller-Nicely data, equal intelligibility is obtained for speech filtered to 200-1500 Hz and speech filtered to 1500-5000 Hz.

Another interesting general result obtained from the Miller-Nicely data is seen in Fig. 3. Here the bandwidth is kept fixed (200-6500 Hz) and the signal to noise ratio is varied. The curve is quite linear (on a db abscissa) until zero db at which point it begins to saturate. The 12 db point corresponds to 3634 correct labels compared to 366 mistakes and this yields 90.8% correct.

TABLE II

17 CONDITIONS FOR MILLER-NICELY CONFUSION MATRIX DATA

	S/N	Filter Band
1	-18	200-6500
2	-12	
3	-6	
4	0	
5	6	
6	12	
7	12	200-300
8	12	200-400
9	12	200-600
10	12	200-1200
11	12	200-2500
12	12	200-5000
13	12	1000-5000
14	12	2000-5000
15	12	2500-5000
16	12	3000-5000
17	12	4500-5000

INTELLIGIBILITY vs. CUT-OFF FREQUENCY FOR "k"

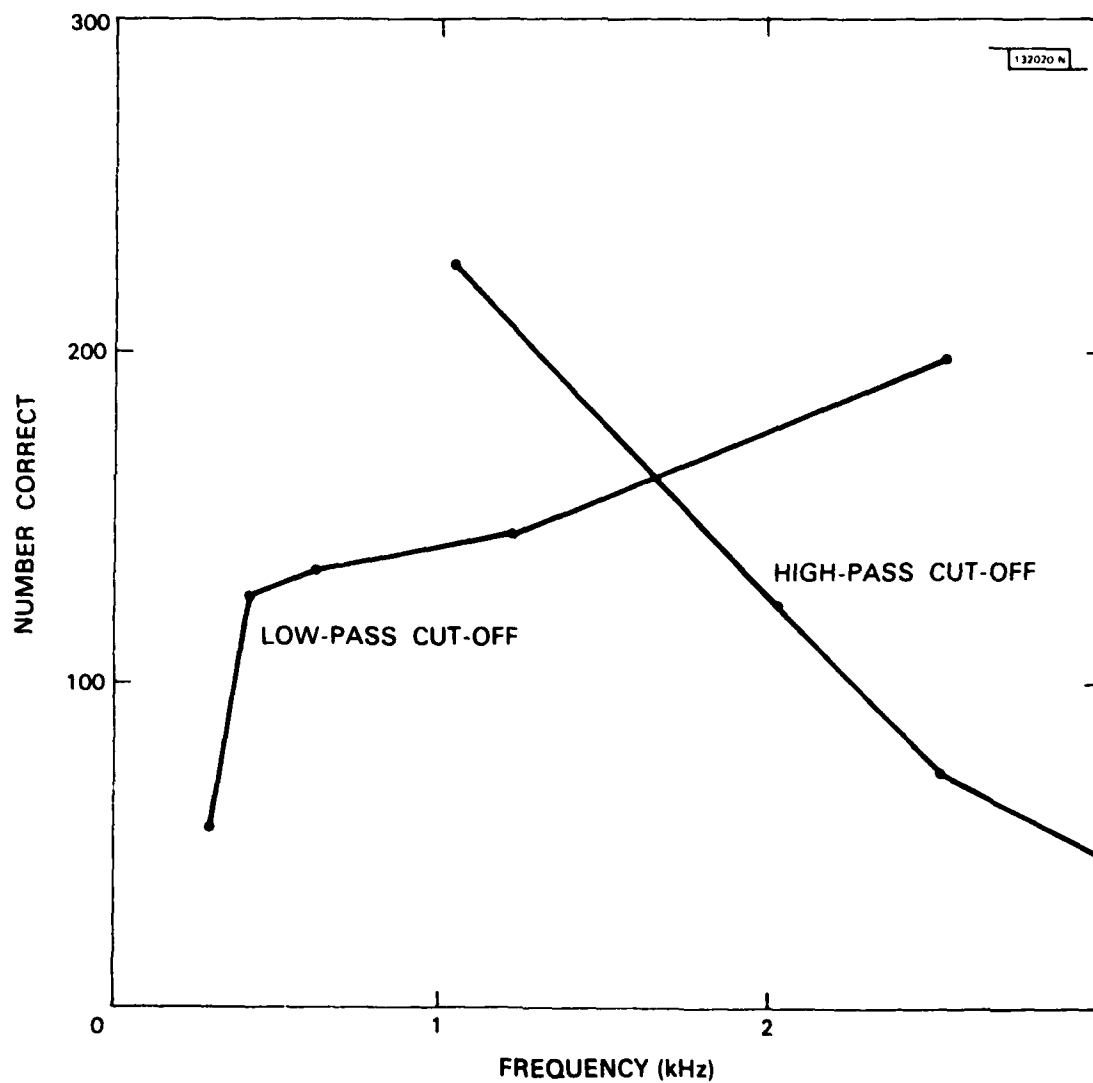


Fig. 1. Intelligibility vs. cut-off frequency for "k" (from Miller-Nicely data).

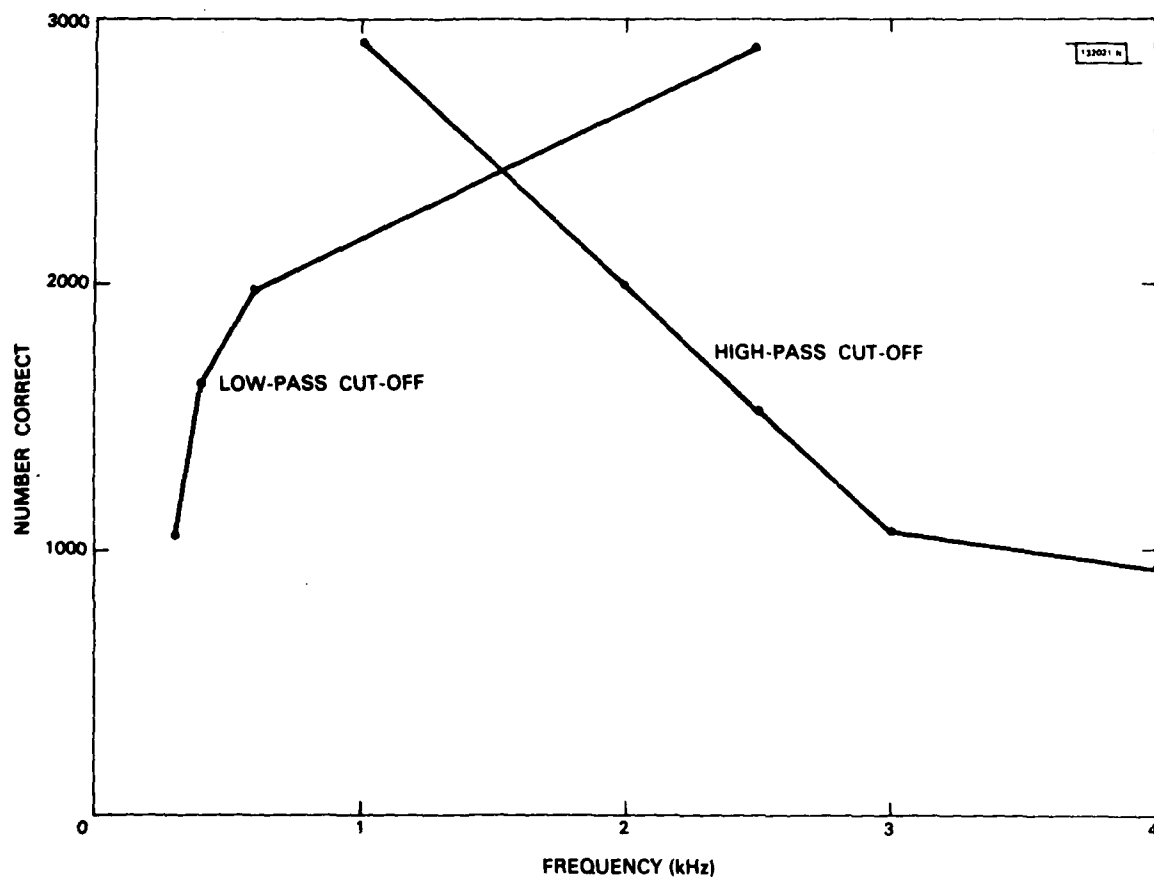


Fig. 2. Intelligibility vs. cut-off frequency (from Miller-Nicely data), all 16 consonants.

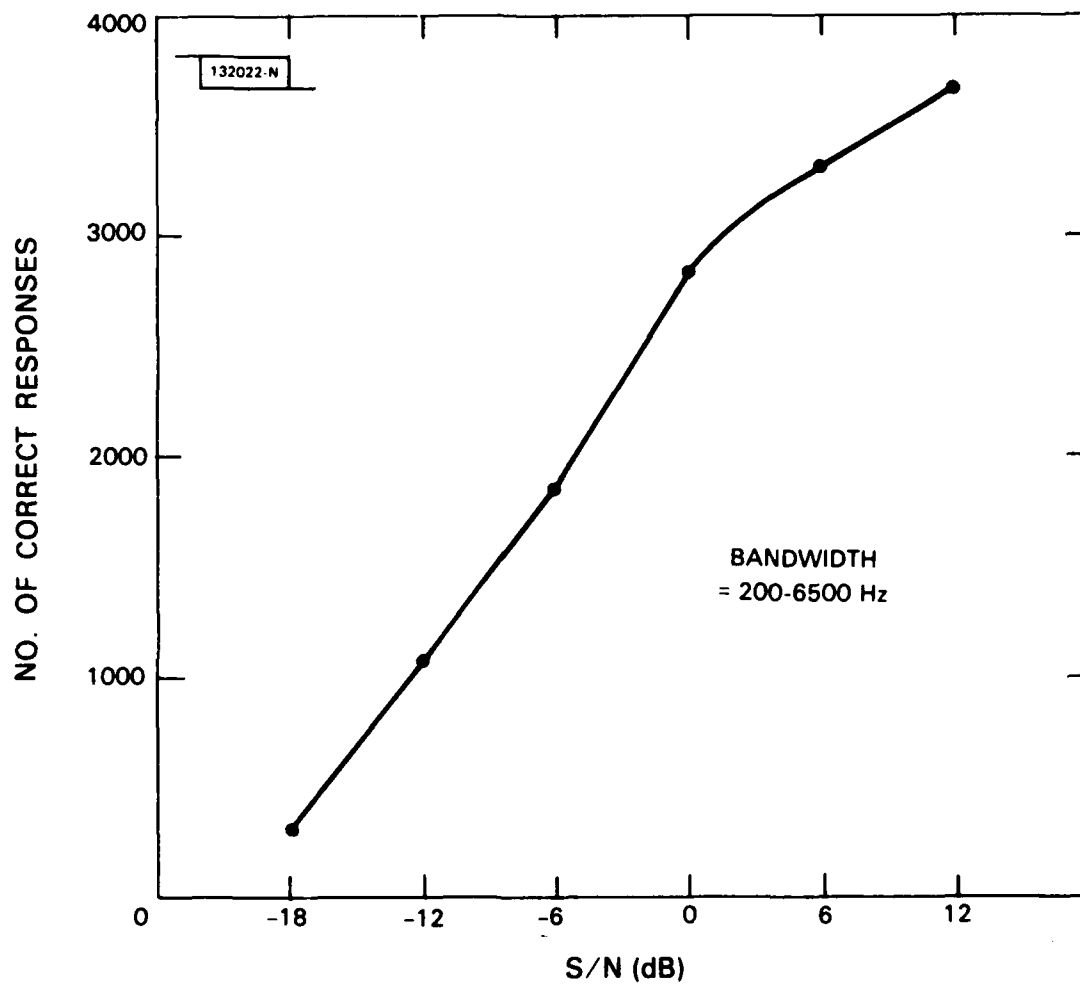


Fig. 3. Total number of correct responses vs. s/n for speech bandwidth 200-6500 Hz (from Miller-Nicely).

Analysis of DRT-Generated Confusion Matrices

For purposes of this section, DRT data was obtained from a variety of vocoder configurations. The recording of the input data and the listener results were done via subcontract to Dynastat. (Appendix B shows examples of confusion matrices.) First, the various configurations listed in Table III will be explained. Then, the DRT scores listed in Table III will be used as a basis for discussing the questions posed at the beginning of this section.

Most of the configurations of Table III are variations of a basic spectrally flattened channel vocoder structure which we have called "flatvoc." This basic structure has the following parameters:

- 10 kHz sampling rate
- 23 channels (filter bandwidths and spacing in Table IV)
- Gold pitch detector [6]
- double filter bank synthesizer
- 100 Hz frame rate
- 3 bit log deltamod coding relative to 6 bit log for low filter
- 8 kbps total bit rate

Three environmental conditions are tabulated. For completeness, some of the results are listed directly from Singer [3]; these are starred.

Pitch vs Spectral Vulnerability

With respect to DRT scores, items 20, 21 and 22 make abundantly clear the relative vulnerability of the pitch and spectral tracks. Item 20 shows that computer-generated noise (corresponding to between 0 and 6 db speech-to-noise ratio) results in a DRT score of 74.2, compared to the noiseless

TABLE III
DIAGNOSTIC RHYME TEST SCORES

Condition: QUIET, DYNAMIC MICROPHONE	JH	PC	RM	Average
1. unprocessed *	98.0	98.6	98.6	98.4
2. 5 KHz RAW:PCM speech with 10 KHz sampling and 12 bit quantization	97.0	97.4	96.1	96.8
3. flatvoc: 8 kbps channel vocoder	94.5	95.6	93.7	94.6
4. Monol50: flatvoc with 150 Hz monotone pitch	92.6	92.2	91.5	92.1
5. M2400: flatvoc with reduced frame rate plus frame fill	87.9	91.5	88.8	89.4
6. Belgard: 2400 bps U.K. channel vocoder				86.5
7. Belgard hiss; belgard with noise excitation				86.4
Condition: fl5	JH	PC	RM	Average
7a. unprocessed:				92.6
7b. unprocessed 5 KHz				90.1
8. flatvoc: 8 kbps * channel vocoder	80.1	87.4	83.3	83.6
9. flatvoc: frame fill *	78.0	85.3	79.4	80.9
10. flatvoc: 4 kbps *	72.9	85.5	75.1	77.9

TABLE III
(continued)

Condition: f15 (continued)	JH	PC	RM	Average
11. base 6: flatvoc with waveform coded sums of outputs of filters 1-6	80.6	86.6	85.4	84.2
12. m2400: flatvoc with reduced frame rate plus frame fill	71.2	83.7	75.1	76.7
13. flathyb12a: flatvoc with uncoded 70-1330 Hz speech inserted	85.8	89.6	86.2	87.2 (0.87)
14. flathyb12b; 1470-4330 Hz	81.9	84.9	85.9	84.2 (0.86)
15. flathyb5a; 70-770 Hz	81.8	85.9	84.0	83.9 (1.11)
16. flathyb15b; 1970-4330 Hz	81.9	85.8	88.2	85.3 (0.66)
17. Belgard: 2400 bps U.K. channel vocoder	68.4	75.5	70.3	71.4
18. Belgard hiss: Belgard with noise excitation	65.6	69.1	64.3	66.33
Condition: ADDED COMPUTER NOISE	JH	PC	RM	Average
19. flatvoc: computer noise = 240	86.6	85.3	81.4	84.4
20. flatvoc: computer noise = 500				74.2
21. clearpitch: flatvoc with computer noise of 500 applied <u>only</u> to spectrum track	74.6	81.1	75.9	77.2
22. clearspec: same as above but noise applied <u>only</u> to pitch track	92.4	93.1	93.5	93.0

TABLE III
(continued)

Condition: ADDED COMPUTER NOISE (continued)	JH	PC	RM	Average
23. wide: cascade of mono150 and flatvoc, with computer noise of 500 added to mono150 output	74.3	78.4	68.6	73.8
24. narrow: same as above <u>except</u> that all analyzer bandpass filter widths are reduced to 40 Hz	55.6	72.9	61.7	63.4

TABLE IV

FILTER BANK FOR 10 KHZ SAMPLING
23 CHANNELS

	3 db Band Edges	Center Frequency	Bandwidth
1.	70-210	140	140
2.	210-350	280	140
3.	350-490	420	140
4.	490-630	560	140
5.	630-770	700	140
6.	770-910	840	140
7.	910-1050	980	140
8.	1050-1190	1120	140
9.	1190-1330	1260	140
10.	1330-1470	1400	140
11.	1470-1610	1540	140
12.	1610-1790	1700	180
13.	1790-1970	1880	180
14.	1970-2150	2060	180
15.	2150-2330	2240	180
16.	2330-2510	2420	180
17.	2510-2770	2640	260
18.	2770-3030	2900	260
19.	3030-3290	3160	260
20.	3290-3550	3420	260
21.	3550-3810	3680	260
22.	3810-4070	3940	260
23.	4070-4330	4200	260

case of 94.6 (item 3). Removal of the noise from only the pitch track (item 21) yields a gain of 3 DRT points. removal of the noise from only the spectral track (item 22) yields a much larger gain of 19.8 DRT points, resulting in a score of 93, which is quite close to the noiseless case.

These results seem to argue for an almost total concentration on spectral (as opposed to pitch) questions, but let us look a bit more. Item 4 actually augments such a viewpoint; it tells us that using a monotone pitch track, the score drops by only 2.5 points (compare to item 3). However, items 6, 7, 17, and 18 tell a somewhat different story. Items 6 and 7 show that, in the absence of background noise, pure noise excitation at the vocoder synthesizer is comparable to a proper pitch track. Items 17 and 18 show that such is no longer the case in the presence of an F-15 acoustic noise background.

More test results are needed before any definitive statements can be made but some preliminary comments are in order. First, it is time to lay to rest the many "folk wisdom" statements that equate lack of widespread vocoder usage with poor pitch detection. The situation is clearly more complicated and if anything, spectral fidelity seems to be the major culprit, although items 17 and 18 indicate that more natural buzz-hiss detection has a beneficial effect in the presence of noise. Second, it appears that DRT scores are not sufficiently sensitive to pitch flaws in the vocoder system. Pitch is a prosodic feature and its real connection to perception should lie in the connected speech rather than isolated word domain.

Aggravation Factor

From items 1 through 6 of Table III we can detect a gradual loss of intelligibility. It seems fair to say that vocoder intelligibility loss is based on the accumulation of small contributions due to several factors. For example, the 1.6 point loss in going from item 1 to item 2 is probably due to the slight bandwidth reduction; the 2.2 point loss from item 2 to item 3 could be due to the spectral sampling and quantization, and (perhaps) the extra reverberation created by the three filter banks of the vocoder. A further loss of 2.5 points (items 3 to 4) is the effect of an unnatural pitch track. A rather substantial loss of 5.2 points (items 3 to 5) is caused by the reduction in effective frame sampling rate in going from an 8 kbps to a 2.4 kbps system. Finally, the 2.9 point loss (items 5 to 6) could be due to a combination of bandwidth reduction and structural differences. Notice that the biggest loss between successive items is only 2.9 points, yet the difference between the reprocessed speech (item 1) and Belgard (item 6) is a substantial 11.9 points.

The most significant fact gleaned from the above is that despite the significant losses, all systems mentioned perform well enough; DRT "folk wisdom" [4] stipulates that DRT scores of 87 or higher can be classified as "good," "very good," or "excellent."

Unfortunately, this state of affairs no longer holds in an acoustically noisy environment, such as an F-15 cockpit. Under these conditions, only items 8 and 11 can be classified (barely) as even moderate, while items 10 and 12 fall in the "poor" class and item 18 is "very poor." These bad results contrast with those of the unprocessed items 7a and 7b, which yield "good" or "very good", despite the F-15 noise.

One might argue that such results are to be expected; after all, acoustic noise strips DRT points away from unprocessed speech and the presence of the vocoder adds sufficient DRT loss to drag the score below an acceptable value. The point, however, is that the vocoders generally aggravate the situation, causing a greater loss in DRT points than one gets by adding the loss due to noise plus the loss due to vocoding. The situation is depicted in Fig. 4. The baseline loss of 5.8 points is simply the difference between items 1 and 7a. The bottom heavy curve shows the results obtained by simply summing the vocoder losses (in the quiet) to the baseline loss of 5.8 points. The top heavy curve shows the actual results obtained from Table III. Note that all systems deteriorate to the next lower class (see Appendix C for a definition of the classes).

The conclusion we must draw is that vocoder processing compounds the pernicious effects due to a strong acoustic noise background; also, those systems that cause more loss in the quiet cause more such compounding. Perhaps this implies that if we manage to improve DRT scores of such basic devices as 2400 bps vocoders, the compounding effect will diminish. To make headway in this area suggests a closer look at the DRT-generated confusion matrices.

DRT Profile of Channel Vocoder Systems

Items 7a through 17 of Table III constitute, more or less, a profile of how DRT scores increase as one goes from 2400 bps systems to the unprocessed speech in relatively small steps. Visualization is improved by plotting scores vs. systems in Fig. 5. As we move to the right, more bits are generally added to the system. Where applicable, the bit rate is

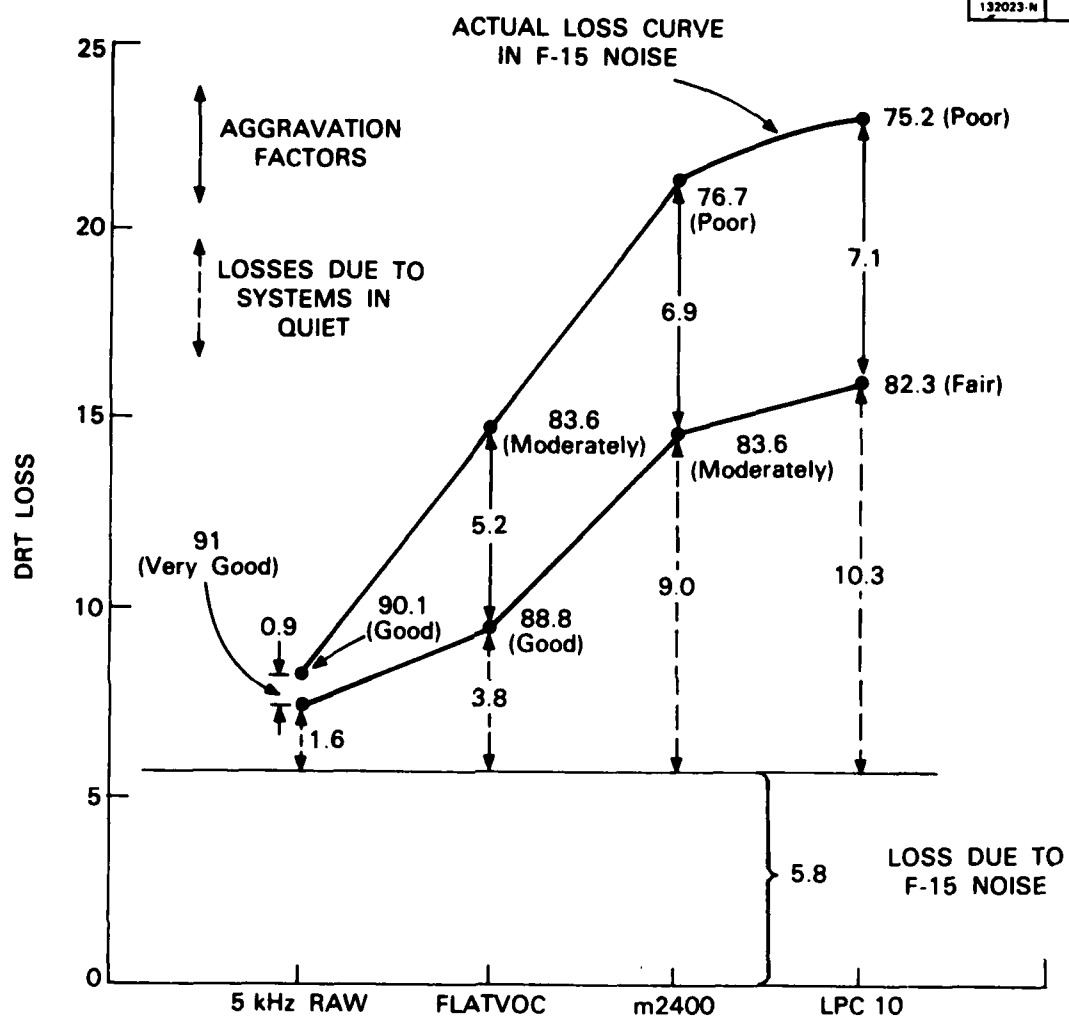


Fig. 4. Aggravation factor for four systems.

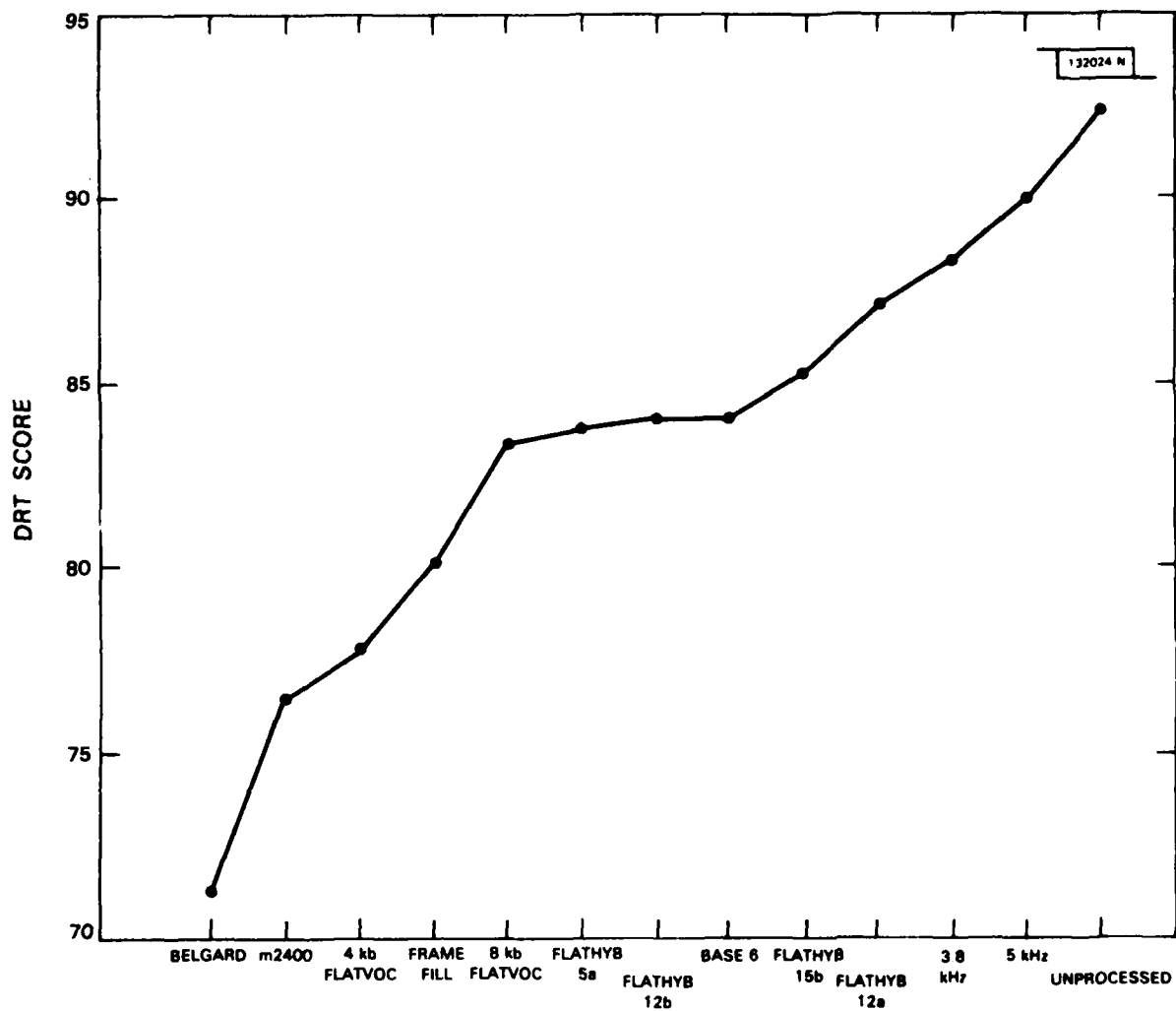


Fig. 5. DRT scores vs. systems in F-15 noise.

shown below each system in Fig. 5. All DRT scores were obtained with the simulated F-15 noise condition.

The single biggest improvement is in going from a strict Belgard structure to the more complex m2400. Both are 2400 bps systems, but m2400 employs spectrum flattening, frame fill and a higher overall bandwidth; all these items, we conjecture, contributed to the DRT jump. When the data rate is increased to 4 kbps, a slight increase is won, but the same 4 kbps system was enhanced even more by the addition of frame fill. A tentative recommendation is that frame fill is a useful addition to any frame-oriented vocoder system.

Another big DRT improvement is won when we move to a straightforward 8 kbps system, but, disappointingly, the addition of a substantial amount of relatively unprocessed speech contributes little to further improvement. In fact, the results obtained from items 14 and 16 seem somewhat anomalous, since item 16 uses less raw speech but achieves a slightly higher score. Also, item 11, which processes the raw speech through the complete spectrally flattened synthesizer does slightly better than item 15, which uses the same bandwidth of raw speech directly.

When we dispense with the vocoder completely, scores consistently improve as bandwidth is increased, as seen in the three right-most items in Fig. 5.

One of our aims is to improve our 2400 bps system to the point where the DRT score is reasonably close to the roughly 90% score achieved by the raw speech in F-15 noise. A reasonable approach would be to divide the problem as follows: a) methods (with a quiet background) for getting

the 2400 bps system to be as good as the 8 kbps system, and b) methods (with an acoustic noise background) for improving the 8 kbps system to where it can achieve close to 90% DRT.

IV. PERCEPTUAL ANALYSIS OF VOCODERS BASED ON PITCH EXPERIMENTS

Introduction

Traditionally, the ear's anatomy has been subdivided into outer ear, middle ear and inner ear. Sound enters the ear and travels down the outer ear until it impinges on the eardrum. This phase can be modelled as an acoustic travelling wave in a fixed cavity. The vibration of the eardrum sets a system of very small bones (the hammer, anvil and stapes) into motion and it is the stapes that is mechanically connected to the inner ear. The inner ear is a quite complicated structure and at this point, it is useful to refer to the very schematized, almost cartoon-like Fig. 6. The stapes mechanically transfers energy to the tympanic fluid in the cochlea and the vibration of this fluid stimulates into vibration a large number of hair cells. These hair cells consist of two types, called the inner and outer. The "hairs", or cilia, on the inner hair cells vibrate mechanically due to the fluid motion. Their mechanical activity modulates the ionic flow across the hair cell membranes. The hair cells synapse on the primary auditory neurons and their electrical activity affects the firing pattern of these neurons. A rough sketch of this progression of influences is shown in Fig. 7. The outer hair cells are connected to the tectorial membrane as seen in Fig. 8 but at present it is not known what role, if any, they play in auditory processing. It is estimated that 95% of the primary auditory neurons are synapsed by the inner hair cells, even

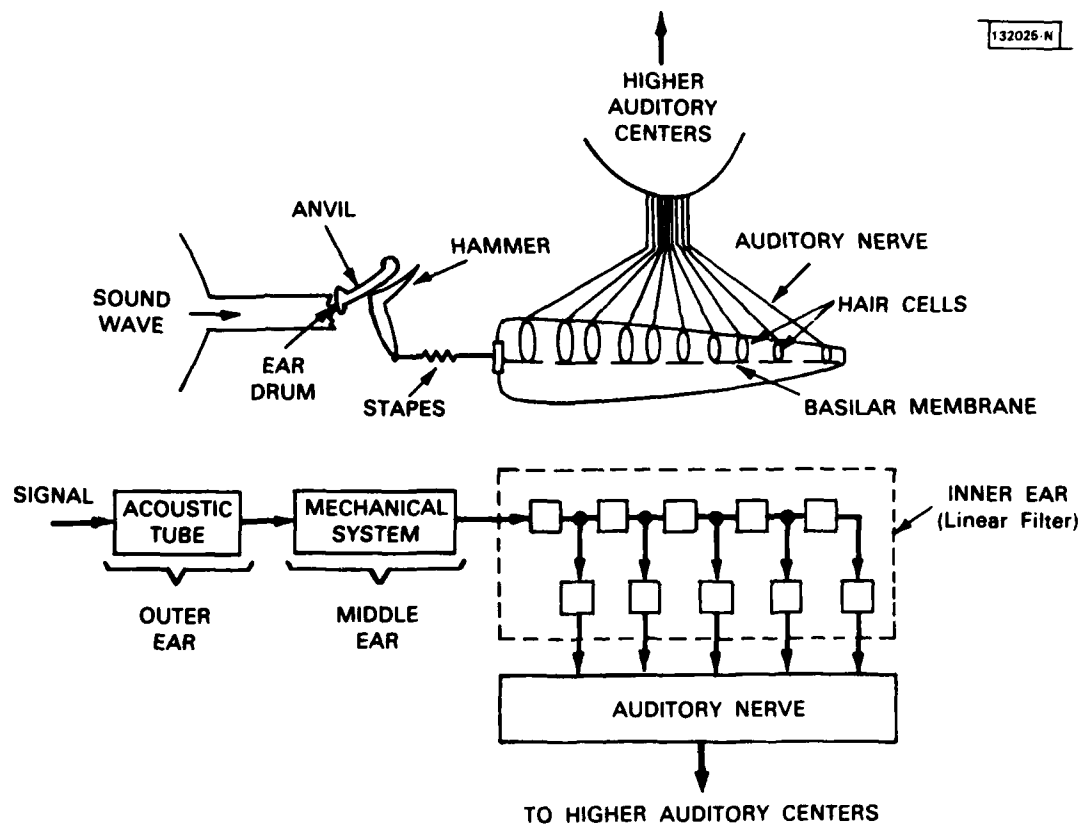


Fig. 6. Cartoon-like representation of the peripheral auditory system.

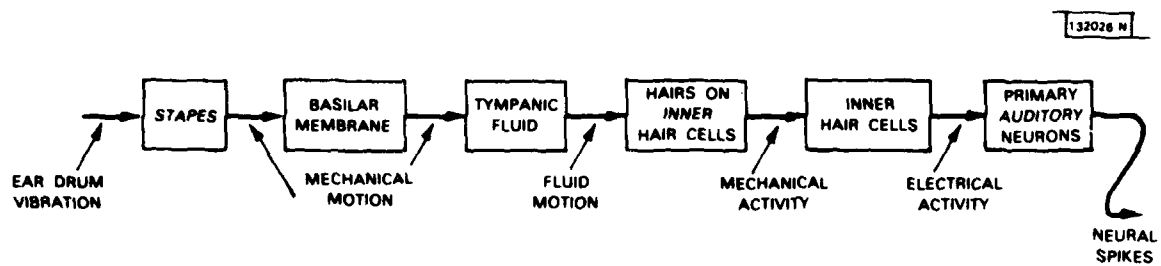


Fig. 7. Block diagram of energy transfers in the ear.

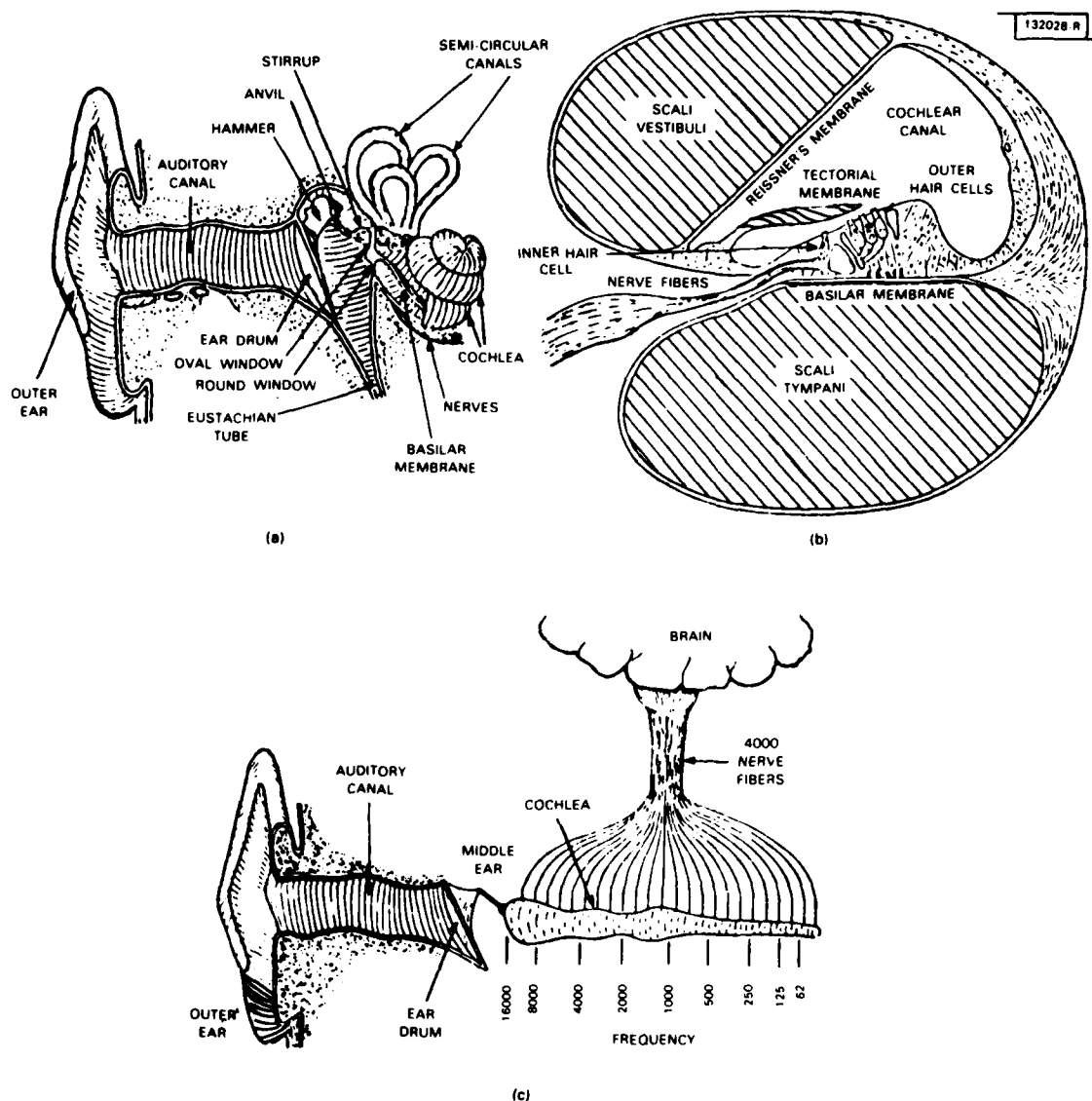


Fig. 8. Some views of the auditory system. (a) Sectional and perspective views of the human hearing mechanism. (b) Sectional view of the cochlea. (c) Schematic view of the human hearing mechanism showing the outer ear, the middle ear, the cochlea, and the nerve fibers leading to the brain.

though there are three times as many outer hair cells. A more elaborate discussion of auditory anatomy and physiology is reserved for Section V.

Helmholtz conceived of the auditory system as a bank of many overlapping band pass filters. Near the entrance to the cochlea, the membrane and associated hair cells respond to high frequencies and as you move to the right on Fig. 6, the response becomes more sluggish, corresponding to filters with lower center frequencies. Thus, for example, a pure tone would cause only a specific place on the basilar membrane to vibrate and this would lead, via the auditory system, to perception of the tone. Tones of different frequencies stimulate different places on the membranes. From an engineering point of view, this model corresponds to a filter bank covering the audio range. By determining which filters contain energy, the pitch of the tone is determined.

But now a complication is introduced when dealing with the perception of complex periodic signals consisting of many harmonics. Even when the fundamental frequency is missing, the perceived pitch is usually that of the fundamental frequency. Fletcher [7] has referred to this result as "the missing fundamental" while Schouten [8] has named it the "residue". Results of this sort lead to theories of human pitch perception as being influenced by the actual period of the signal rather than by a place mechanism. For the perception of the missing fundamental to be explained by a place theory would require that some non-linear phenomenon in the ear would cause the place in the basilar membrane that corresponds to the fundamental frequency to vibrate despite the physical absence of the fundamental. Licklider [9] proved that this did not happen via the

following clever experiment: he alternately played a pure tone (at the fundamental) and a harmonic series of the same fundamental frequency but with the actual fundamental physically absent. The listener then perceived two sounds of equal pitch but different timbre. Then noise was added to the above sequence; the pure tone was completely masked while the harmonic series came through loud and clear. If perception of the harmonic series were dependent on combination tones appearing at the fundamental frequency place in the basilar membrane, it too would have been masked out.

Licklider's experiment cast a vote in favor of Schouten's (and Seebeck's [10]) theory of periodicity over place.

Further insight was obtained from the relatively recent experiments of Houtsma and Goldstein [11]. In the first experiment, musically trained subjects were asked to recognize "intervals" between two successive signals. Each signal contained two successive harmonics of a given fundamental frequency. The "interval" is defined as the difference between successive fundamental frequencies. When the above two signals were presented to both ears, the trained subjects had no trouble identifying the intervals.

The second experiment was a repeat of the first experiment but with one notable exception; for each signal, one harmonic only was presented to one ear while the other harmonic was presented to the other ear. Again, pitch intervals were correctly identified. This indicated that the perception of pitch was centrally located, that is, it took place after the auditory signals from the two ears had been combined. What actually happens physiologically is still a mystery, but these experiments must now

be incorporated into any proposed model. Also very significant is the fact that place theory (somewhat altered) sneaks back in. We can imagine that the appropriate places on the basilar membrane vibrate for all existing harmonics of the signal and then that the "central processor" combines this knowledge to produce a pitch value.

A straightforward interpretation of the above is that the ear and brain perform a high resolution spectrum analysis followed by a pattern recognition procedure to detect pitch based on this spectral structure. Goldstein [12] has proposed a model of this sort, and Duifhuis et al. [13], have implemented such a model to extract pitch from a speech wave.

Psychophysics Experiments That Must Be Explained by a Pitch Perception Model

If we are to stand by our original hypothesis that the auditory system is an effective pitch detector, then any proposed model should yield results that agree with known psychoacoustic data. Following is a brief summary of some of this data:

a) Any model must be able to extract the pitch from a periodic signal with missing fundamental.

b) De Boer [14] describes some experiments on inharmonic signals. He first generated the frequencies 1400, 1600, 1800, 2000, 2200, 2400, 2600 Hz and established the pitch to be 200 Hz. Then he shifted each harmonic by a fixed amount (say 30 Hz) to obtain 1430, 1630, 1830, 2030, 2230, 2430, 2630 Hz. Although this sequence results in an inharmonic signal, the perceived pitch is about 205 Hz, illustrating that approximately periodic signals evoke the perception of the approximate pitch. This should not be too

surprising when we consider that the pitch of a speech wave is fairly well tracked despite its variation.

c) Miller and Taylor [15] have evoked pitch perception by switching white noise on and off periodically. Since the spectrum of this signal remains white, it is difficult if not impossible to invoke a place model to explain this experiment.

d) When two harmonically related sine waves are played in sequence, the pitch is perceived to be that of each tone [16]. When noise is added, the fundamental frequency is perceived. This experiment argues for some sort of noise-dependent smoothing by the auditory system.

The Principle of Dominance

Although human pitch perception can be based on either a low passed or high passed version of an audio spectrum, Ritsma [17] has proved that the low frequency portion of the spectrum is dominant. He showed this with a simple but ingenious experiment wherein he added the outputs of a low pass and a high pass filter. Both filters were excited by periodic pulse trains of different fundamental frequencies; Ritsma's subjects responded most strongly to the low passed version. Our informal experience with vocoders verifies this principle of dominance for speech. In a hybrid vocoder consisting of vocoded speech added to waveform-coded (baseband) speech, the baseband should be at the low end of the spectrum to minimize the perception of pitch errors. It is also true that the most effective portion of the spectrum for performing automatic pitch extraction is in the baseband region (200-1000 Hz).

The analysis of Section III indicates quite strongly that DRT scores are more severely affected by noise in the spectrum channel rather than noise in the pitch channel. On the other hand, we know that the breakdown of the pitch track can cause the vocoded speech to be unacceptable. Furthermore, although much notable pitch perception research has been done using non-speech stimuli, little such work has been done with human (not synthetic) speech. For these reasons, we decided to examine in some detail how speech degradations and vocoder structures influence the pitch percepts of vocoded speech. At present, we are implementing the facilities needed to perform this work and performing informal pilot experiments to obtain a "feel" for the kind of data to accumulate. Thus, this section is a preliminary statement of progress toward the goal of more formal pitch perception experiments using speech stimuli.

Ritsma's "dominance" principle suggests that pitch is most strongly perceived in the spectral region 200-2000Hz. From this, we expect that pitch and buzz-hiss errors are most noticeable in that region. To test this, we ran an informal experiment where we combined low pass filtered speech with noise-excited vocoded speech. The subjects had control of the cut-off frequency of the low pass filter with the vocoded speech spectrum always remaining contiguous. Although differences between this hybrid signal and the full band speech signal could be detected at cut-off frequencies above 2 KHz, the consensus was that a cut-off frequency of about 1500 Hz was sufficient to create acceptable hybrid speech; the main effect appeared to be a more pronounced "breathiness" of the speakers.

In a vocoder, pitch is sampled and quantized and perhaps smoothed. This raises the question of the correct degree of these processes and this in turn hinges on the ability of the listener to discriminate. To examine these issues, a pitch discrimination experiment is presently being arranged. Well-known results exist for pitch discrimination of pure tones and pulse trains, and we intend to extend such results to encompass the spectrum of pitch-discriminable signals from pure tones to vocoded speech, including constant frequency pulse trains, pitch tracks, steady-state vowels and monotone speech. A rough experiment was run on the discriminability of vocoded speech from our 8 kbps channel vocoder. Tape S1¹ was played through the system. The first of each sentence pair was played through the vocoder with normal pitch; for the second sentence, the overall pitch contour was lowered or raised by a given (random) percentage. It was found that the threshold of discrimination for lowered pitch was about 3% and for raised pitch about 2%. Tests using the various stimuli mentioned above will be conducted more formally.

Further insight into dominant regions can be obtained via masking experiments. If the low frequencies are dominant for human pitch perception, it should follow that additive low frequency noise should cause greater deterioration of perception than the same amount of high frequency noise. Some informal listening indicates this to be true, but formal testing is required before any reasonable quantitative statements can be made.

Returning to the question of whether spectrum or pitch track noise creates more listener difficulties, recall the results of items 20, 21

¹"S1" refers to the standard tape generated by the Narrow Band Speech Consortium (1975).

and 22. Item 21 says that 3 DRT points are gained by removing the noise from the pitch track while item 22 tells us that 18.8 points are gained by removing the noise from the spectrum track. Thus, one might conjecture that in an operational vocoder system, as noise is increased, spectrum information gets clobbered much before pitch information. However, results for items 20, 21 and 22 were obtained using computer-generated noise with an approximately white spectrum. In a sense, this biases the results toward favoring pitch, since: a) we have seen that pitch perception is dominant in the low frequency region (i.e., 200-1000 Hz), and b) the Gold pitch detector, used in all experiments, examines only this low frequency region. Thus, if white noise is present at the vocoder input, the actual signal-to-noise ratio at the input to the pitch algorithm (after low pass filtering) is approximately 12 dB higher than the corresponding S/N at the spectrum input.

The "fair" approach to this problem is as follows. We first collect pure noise tapes from different aircraft (for example, F-15, helicopter, E3A) and add this noise to any part of the vocoder. Thus, for example, this noise can be added to either the pitch or spectral track separately or even to a portion of the spectral track. Furthermore, the noise levels at different inputs can be independently adjusted. We are in the process of building the hardware and software needed to run such experiments.

Hall and Peters [16] have found that when harmonics of a fundamental frequency are presented time-sequentially, subjects can identify the individual harmonics, but when noise is present, the tones "merge" and a percept of the fundamental frequency is obtained. This implies that

subjects perform smoothing over longer time intervals in the presence of noise. Such a result may have a pay-off in vocoders that are designed to work in noisy environments. One can imagine imposing greater smoothing on both pitch and spectrum tracks during noise. At present, these notions are still in the "idea" stage, but we hope to exploit them in the near future.

V. THE PERIPHERAL AUDITORY SYSTEM

Introduction

If we understood completely the behavior of the auditory system in the human - from the periphery system consisting of the ear right up to the auditory portion of the cerebral cortex - it would be relatively easy to design speech transmission systems. We would only need to simulate this system on a digital computer, present an acoustic signal to the input and observe which features the system extracts from the acoustic input. Having made this measurement, we would then present to this simulation the output from any communications system under design and make sure that the same undistorted features were present at the output of the simulation as when the unprocessed speech was presented. Additional knowledge about the range of distortions allowed on speech features of importance to the perception process would allow us some engineering bounds on systems under design and consideration. This would indeed be a very orderly process wherein the creative part of the design would consist of inventing new techniques for communicating, rather than being concerned about the design's acceptance to the human ear. The acceptance would be easily quantified using the simulation and knowledge of the important features in the perception process.

At the present time, we do not have a very good understanding of the auditory system in the human. From a physiological approach, it is difficult to study the human auditory system in living systems. On the other hand, much was learned by von Békésy and others using ear preparations from cadavers. This detailed anatomical knowledge has been invaluable. To study the hearing process in live animals, it has been necessary to move down the evolutionary chain and find animals that have similar, or what is believed to be similar, auditory systems at least at the level under study. This approach leads to the study of cats, squirrel monkeys, lizards, turtles, and other creatures of opportunity for the physiologist. What we know of the human auditory system has been inferred and hypothesized from a body of ingenious psychoacoustic experiments which have gathered data on how humans perceive and detect certain acoustic stimuli. This combination of physiological studies on animals (and humans under very special conditions of trauma or disease) along with psychoacoustic studies on humans has led to a large body of knowledge on the audition process, but much more research is required before we can fully explain how speech is converted from a series of acoustic phenomena to perceived meaning. In this section, we will outline the physiology of the peripheral auditory system in the human and point out in parallel what the implications are for speech communications and compression equipment designs.

The Peripheral System

The peripheral auditory system includes the outer, middle, and inner ear, as well as the cochlear nerve bundle which connects the cochlea or

inner ear to the brain. Figure 9 shows a cross sectional view of the system and the block diagram indicates the nature of each portion. The outer or external ear provides a mechanical connection to the acoustic signal in free space. This acoustic signal is then coupled to the middle ear apparatus through the eardrum, which is connected to a series of three small bones - hammer, anvil, and stirrup - and then in turn to the inner ear. This chain of action which transmits the incident acoustic signal to the inner ear is linear for the most part. The movement of the stirrup or stapes is coupled to the cochlea or inner ear by means of a membrane called the oval window that vibrates with the stapes and generates fluid waves in the fluid filled cavities of the cochlea. Inside the cochlea, two structures cause a conversion of the acoustic fluid wave into a coded signal which moves along the cochlear nerve to the brain. Before we look in more detail at the inner ear transducer mechanisms, we should note the feedback paths from the nervous system back to outer, middle, and inner ear. All of these paths serve to orient and gain adjust the incoming signal. In the case of the pinna or outer ear, the feedback serves to orient the organ for maximum reception of the signal. This does not happen in man, but does take place in most animals. Feedback to the middle ear serves to adjust the middle ear transfer function gain so as to keep the inner ear from being overdriven by loud inputs. Finally feedback to the inner ear organs is less well understood, but appears to serve as an automatic gain control mechanism at the nerve cell level of operation.

Cochlea and Basilar Membrane

The inner ear is a spiral shaped organ that can be imagined as stretched out into a single long fluid-filled set of cavities whose cross

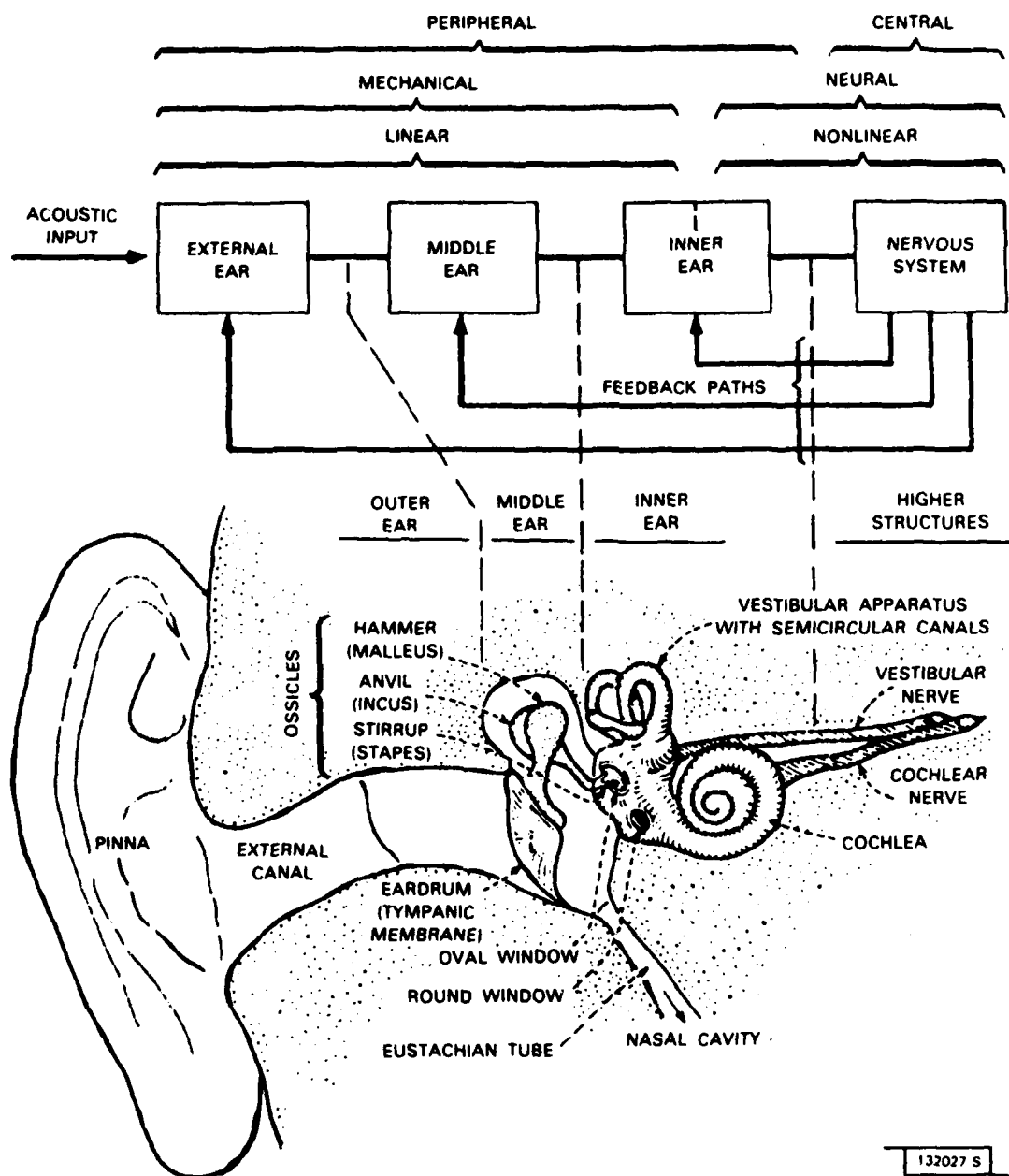


Fig. 9. Block diagram of peripheral auditory system. From Weiner (1949).

section is as shown in Fig. 10. The top and bottom cavities, the scala vestibuli and scala tympani, respectively are connected at the far or apical end (the end away from the stapes) by a hole in the membrane separating them called the helicotrema. The stapes driving the oval window membrane cause fluid waves to be set up in the continuous cavity formed by the scala vestibuli and scala tympani. Separating these two cavities except at the helicotrema is a membrane which vibrates with the fluid waves and effects the frequency analysis of the acoustic signal necessary for the hearing process. This membrane is actually a complicated set of structures of which the most important is the basilar membrane as shown in Fig. 11. This figure presents the two cavities we have discussed, but now we see them in cross section as well as seeing that the dividing partition is actually a complicated structure in its own right which includes another smaller cavity, the scala media. The lower dividing surface between the scala tympani and media includes the basilar membrane which vibrates up and down in this cross sectional dimension when fluid waves occur in the cochlea, and also includes the transducer organ from mechanical to nerve response, the hair cells. The basilar membrane structure is tuned in such a way that the end closest to the stapes and oval window, the basal end, responds best to high frequencies, and the farthest end responds best to low frequencies. The response to a single acoustic pulse at the ear drum is shown with place along the membrane as a parameter in a simulation by Flanagan (Fig. 12). Notice that there is a delay for the lower frequency responses since they occur farther down on the length of the membrane. Also although it is not obvious from the figure, the membrane is tuned as a

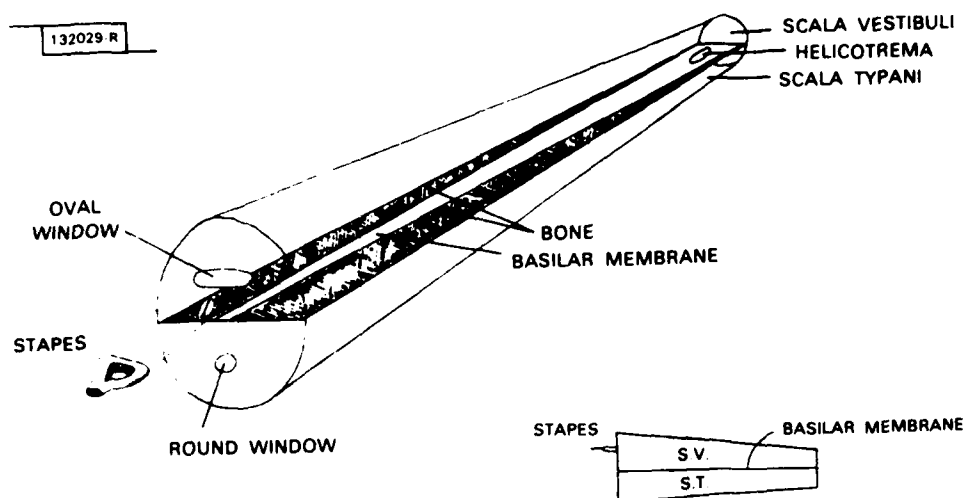


Fig. 10. The major structural features of the uncoiled cochlea. Note that the basilar membrane is narrow near the round window and wider near the helicotrema, a taper opposite the cross-section area of the cochlea.

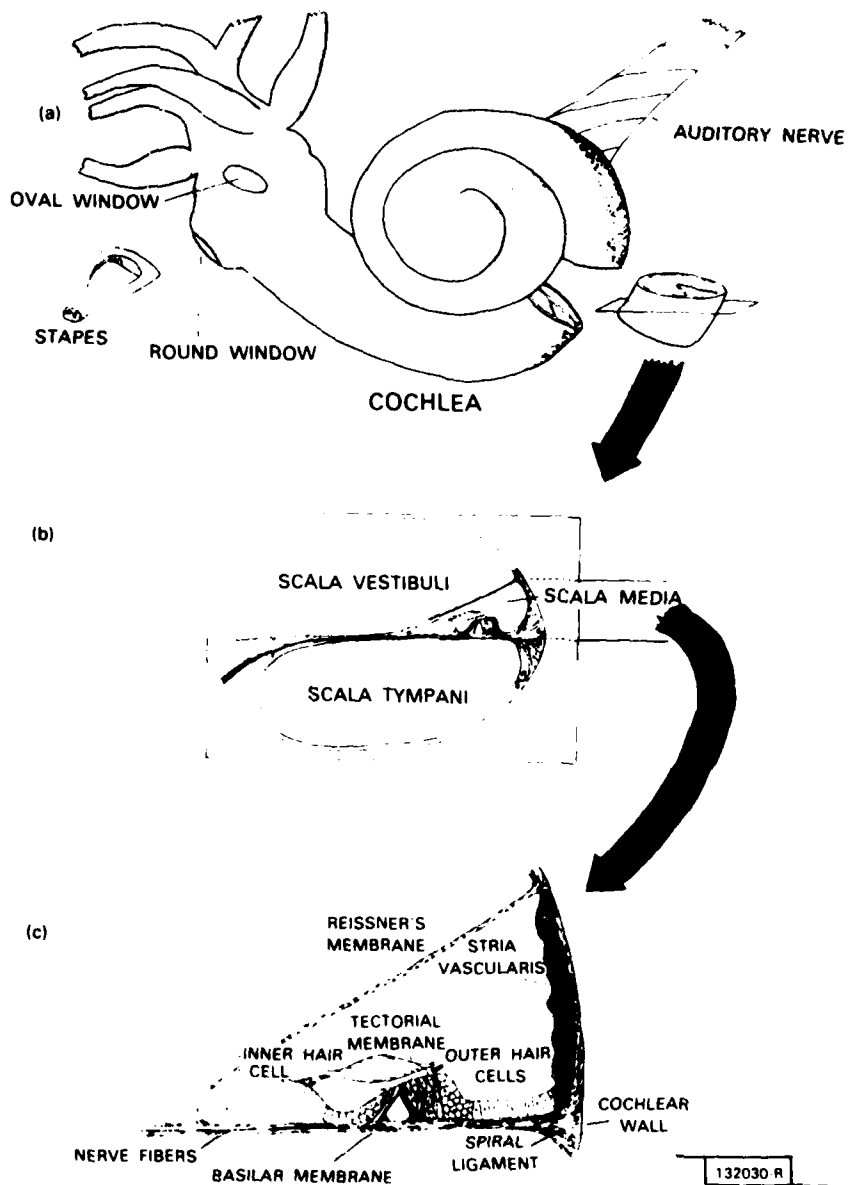


Fig. 11. Structural and anatomical features of the cochlea. (a) The cochlea in relation to the middle ear and auditory nerve. (b) Cross section of the cochlea. (c) The scala media (from Green "An Introduction to Hearing").

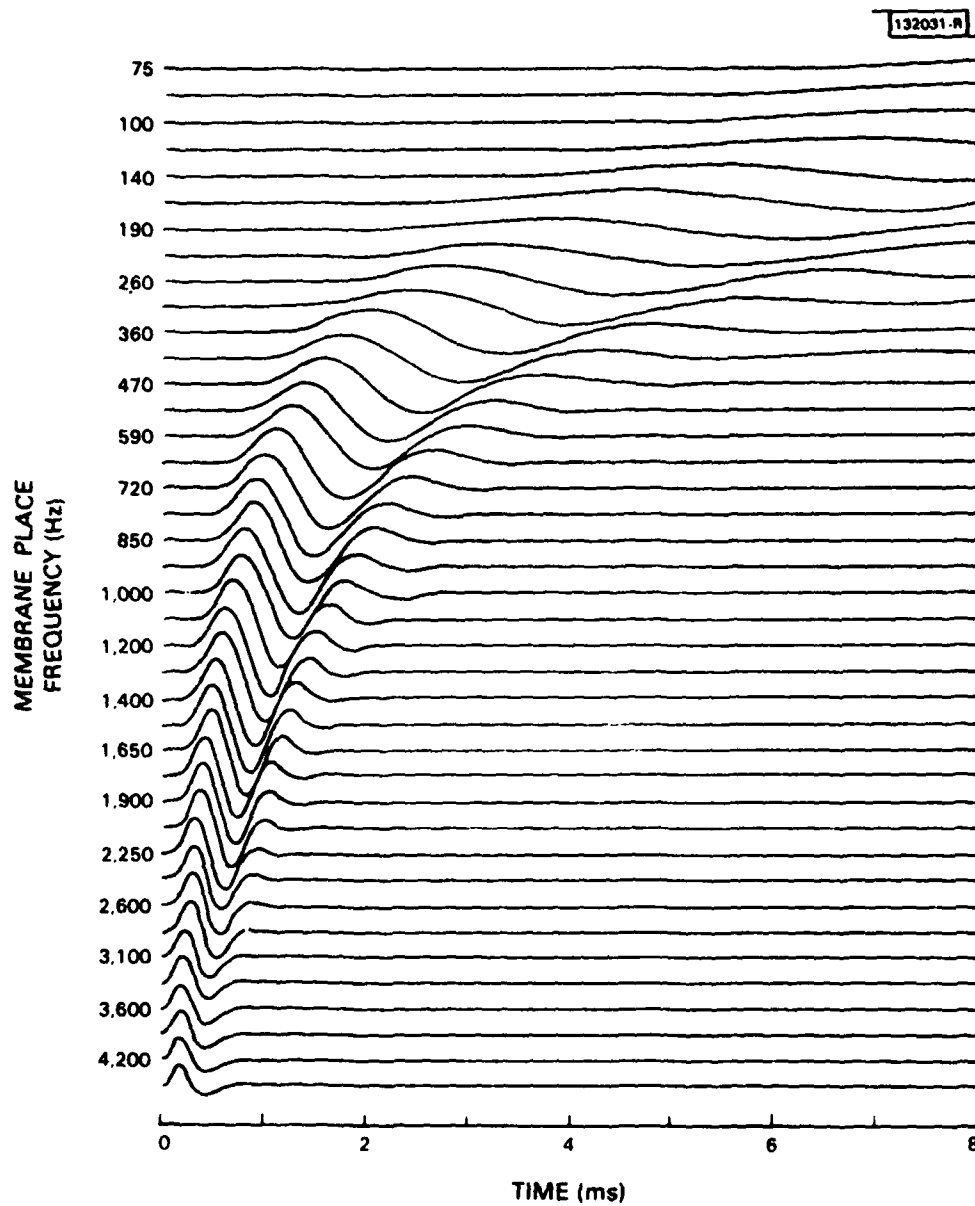


Fig. 12. Computer model of basilar-membrane displacement. Response to a single rarefaction impulse of sound pressure at the eardrum.

function of position in almost a logarithmic manner so that logarithmic distance along the membrane corresponds to linear frequency change. Notice also from the figure that the higher frequency responses have more resolution indicating an increase in bandwidth with frequency. The individual place responses are shown in normalized form in Fig. 13 which displays both magnitude and phase response versus frequency for a normalized filter. Notice the sharper cutoff response on the high basilar membrane portion of the cochlea; this leads us to several conjectures about auditory processing and its implications for speech communications equipment. The nature of the basilar membrane "filters" indicates that speech communications devices must have good resolution in the time domain for high frequencies. This follows from the properties of the membrane we have just described. We should note that no vocoders of either the LPC or channel variety use this property in any meaningful way. The LPC vocoder approach uses a uniform bandwidth analysis across frequency which is a function of the time window used. The channel vocoder at best uses a filter bank for channel analysis which may increase bandwidth as center frequency increases. The detected envelopes of these filter outputs are all smoothed with the same bandwidth low pass filter (usually around 25-30 Hz), thus reducing any time resolution associated with the higher bandwidth higher frequency filters. There are techniques which can incorporate these suggestions into both the LPC and channel vocoder framework, but they have not yet been explored. Consequently we do not know if this extrapolation from our knowledge of basilar membrane behavior will add to speech communications quality and intelligibility.

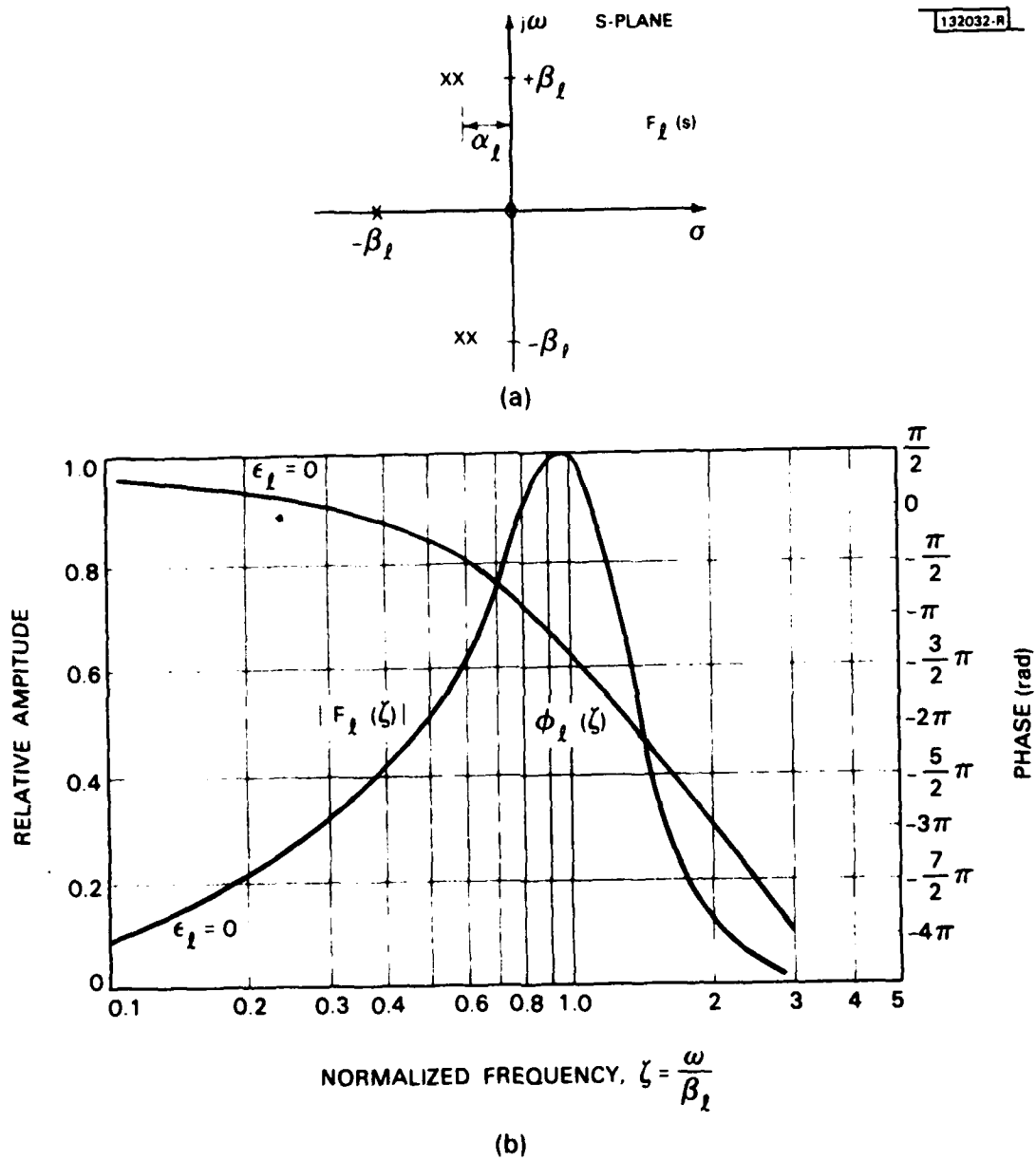


Fig. 13. Analytical model for basilar-membrane displacement. (After Flanagan, 1962).

Hair Cell Properties

An enlarged view of the central section of the cochlea displays the organization of cells around the basilar membrane. In particular, in Fig. 14 we can see the structures called external and internal hair cells which span the space between basilar and tectorial membranes. It is these hair cells that transduce the mechanical motion of the basilar membrane into electrical activity of nerve cell fibers connected to them. The hair cells in the human inner ear, as in higher animals such as the cat, are organized into three rows of external hair cells, and one row of internal hair cells as one moves along the long dimension of the cochlea. As the basilar membrane moves up and down under the influence of fluid waves, the hair, and in some cases the cell bodies, are subject to distorting forces. The resultant physical changes generate time varying voltages in the hair cell bodies. These time varying voltages interact with the nerve fibers connected to the hair cells so as to cause the nerve cells to fire in patterns correlated to the time varying voltages. Figure 15 is a representation of the movement and distortions associated with the internal (inner) and external (outer) hair cell columns when there is movement of the basilar membrane. Note that the three rows of external cells are fastened between the basilar membrane and the tectorial membrane as shown in Fig. 15, but the inner hair cells are not fastened to the tectorial membrane. It is conjectured that outer cells are sensitive to the basilar membrane position, while the inner cells are sensitive to fluid motion in the scala media.

In the human ear, there are about 4000 inner hair cells arranged along the length of the basilar membrane along with 12,000 outer hair cells which

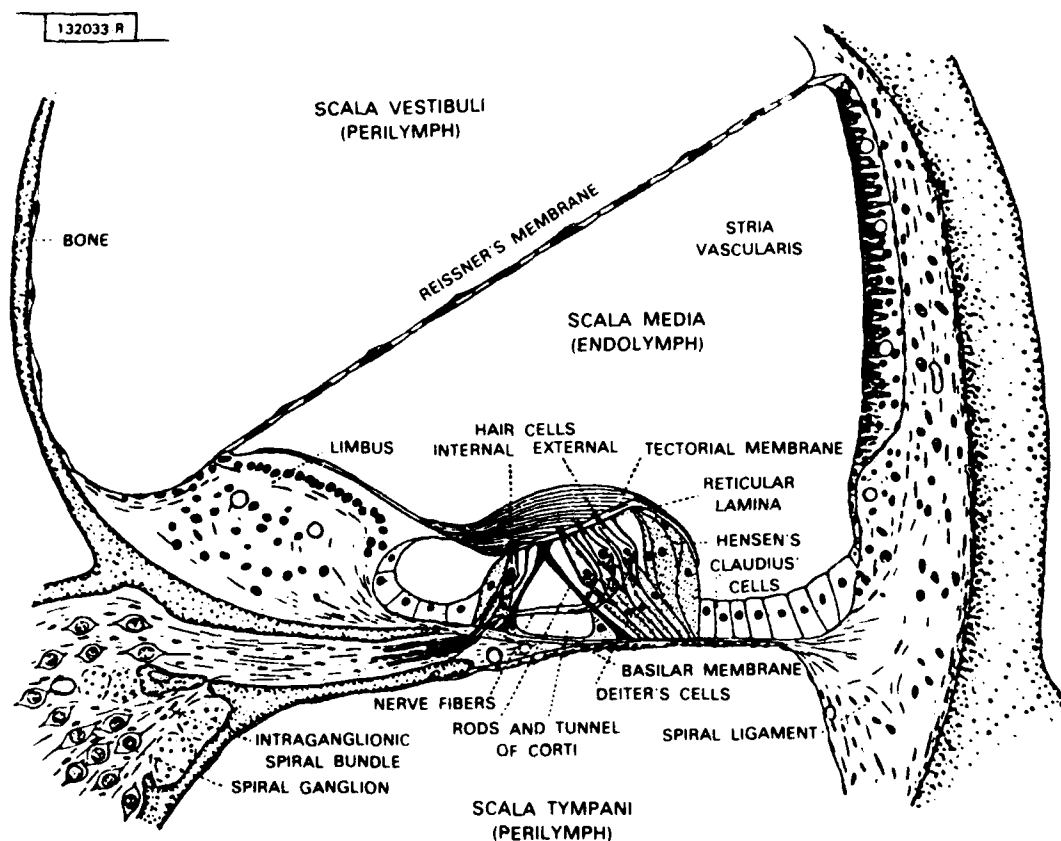


Fig. 14. Camera lucida drawing of a cross section of the cochlea partition in the second turn of a guinea pig cochlea. The attachment shown here of the tectorial membrane to the inner supporting cell, and to Hensen's cells, is based on microdissection of fresh, unfixed specimens. From Davis (1961).

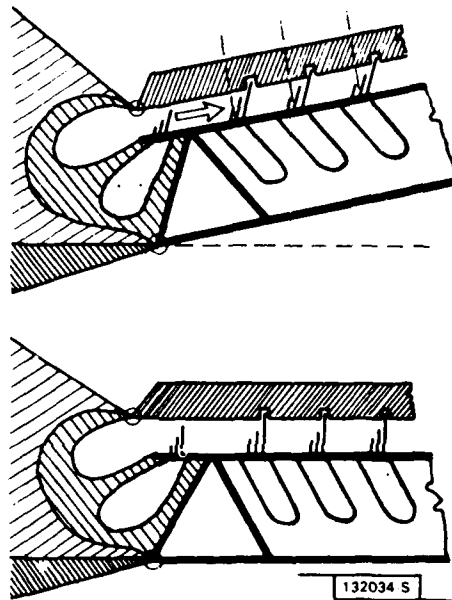


Fig. 15. Relationship between the tectorial membrane and cilia of outer hair cells. At rest (lower illustration) the cilia stand perpendicular to the cuticular surface of the cell. When pressure waves move the basilar membrane, a "shearing" force acts to alter the angle of the cilia with respect to the cuticular surface. Note that the cilia of the inner hair cells are shown to bend, not from tectorial membrane attachment but from fluid motion. From P. Dallos and A. Ryan, : Physiology of the inner ear. In J. L. Northern (ed).: Hearing Disorders, 1976, p. 95, (Little Brown and Co.).

compose the three outer rows. Intertwined with and innervating these 16,000 hair cells are approximately 30,000 nerve fibers which come together to form the eighth or cochlear nerve. The inner hair cells are more primitive cells in an evolutionary sense, but are innervated by about 95% of the 30,000 fibers. The three rows of outer cells are involved with the remaining 5% of the fibers. Although studies on hair cells in the turtle and lizard indicate hair cells to possess an internal tuning mechanism in addition to any tuning in the basilar membrane, there is no concrete evidence for this in the cat or the human. As a consequence, it is generally assumed that the basilar membrane tuning characteristics determine the major tuning in the auditory system.

Since the basilar membrane is a continuous mechanical structure, we can think of the frequency analysis as performed by a bank of highly overlapped (in frequency) filters. In addition, since there are some 30,000 nerve fibers connecting these "filters" to the brain by connecting to hair cells in many redundant and overlapped ways, information flowing to the brain must represent a highly overlapped filter bank analysis with many redundancies in the representation of each filter. Perhaps some of the redundancy provides for dynamic range and amplitude coding in some fashion. It is interesting to speculate that this redundant and overlapped spectrum representation must in some way enhance the processing of noisy speech signals. In addition, if we consider a particular frequency range, the many redundant and overlapped channels coded by the nerve fibers represent a statistical sample over space that may be equivalent to a sample averaged over some time period. This fits with ideas that suggest

the speech signal is coded in some statistical rather than deterministic way when going from the cochlea to the brain. For designers of speech communications equipment, and in particular vocoders, there may be some importance in using many overlapped frequency channels to reduce the effects of additive noise upon vocoded speech analysis.

From Eighth Nerve to Cerebral Auditory Cortex

As we have outlined, an acoustic signal incident upon the outer ear results in a chain of action through basilar membrane vibration and hair cell transduction to produce some manner of correlated nerve fiber pulses or "spikes" of depolarization voltages that travel along the nerve fibers in the cochlear nerve to the brain. The cochlear nerve bundle contains fibers which represent action from all of the frequency spectrum processed by the vibrating basilar membrane. The voltage pulses measured on any one fiber by a microprobe carefully placed into the cochlear nerve are associated with a small set of hair cells and consequently a small area or place on the vibrating membrane. Because the nerve firings are not deterministic when the hair cells excite neurons but instead produce events or firings in a statistical sense, statistical measures must be used to see any interesting data. For example, a common mode of processing data from single fibers is the interval histogram. During excitation with some acoustic probe signal, a histogram of successive intervals between pulses is built up. This measure shows the characteristic period or frequency of the particular fiber and associated hair cells. Figure 16 shows four histograms gathered from four different fibers when the eardrum is excited by a narrow click signal (an impulse-like signal). Notice that the

132035 R

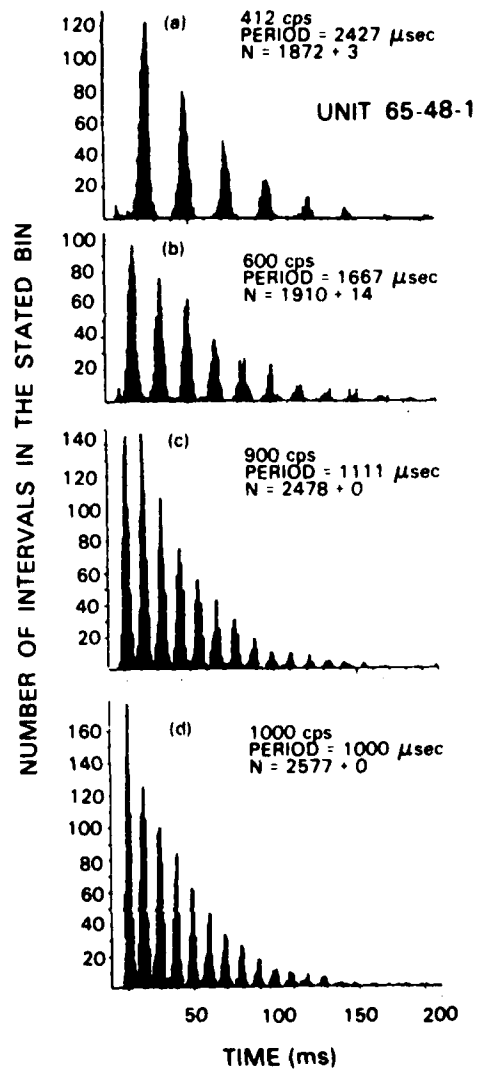


Fig. 16. Interval histograms for tone bursts at different frequencies.

Interval between peaks is related to the characteristic frequency of the particular fiber probed. If the pulse train was examined on a scope display, very little regularity would be seen, but the histogram over enough trials shows the information that must exist if the fiber is transmitting useful data to the brain. Workers such as Delgutte, Sachs and Young, [29,30] have demonstrated that classical parameters of speech such as voicing period and formant frequencies are in evidence from measurements made on these fibers when the ear is excited by artificial vowels that can be repeated and statistical data obtained. This result is, in some sense, to be expected, because we know that speech parameters must be transmitted along the cochlear nerve to the brain, but it has been very difficult to show this information represented on the nerve fibers with real data and real animals.

The auditory pathways from the cochlear nerve to the auditory portions of the cerebral cortex have been explored at least anatomically in many animals, so much information is known about the gross anatomical organization of pathways from inner ear to cortex. Details of physiology and organization at individual assemblages of nerves; nuclei; as they are referred to inside the brain, are very much lacking. A pictorial sketch of some of the higher pathways is presented in Fig. 17. The figure traces the pathway from the organ of corti which contains the basilar membrane and the hair cells to the first group of neurons in the brain receiving these fibers and called the cochlear nucleus. From this nucleus there are pathways of fibers that both cross over to a nucleus associated with the other ear and also ascend to the next nucleus (the superior olive) from the

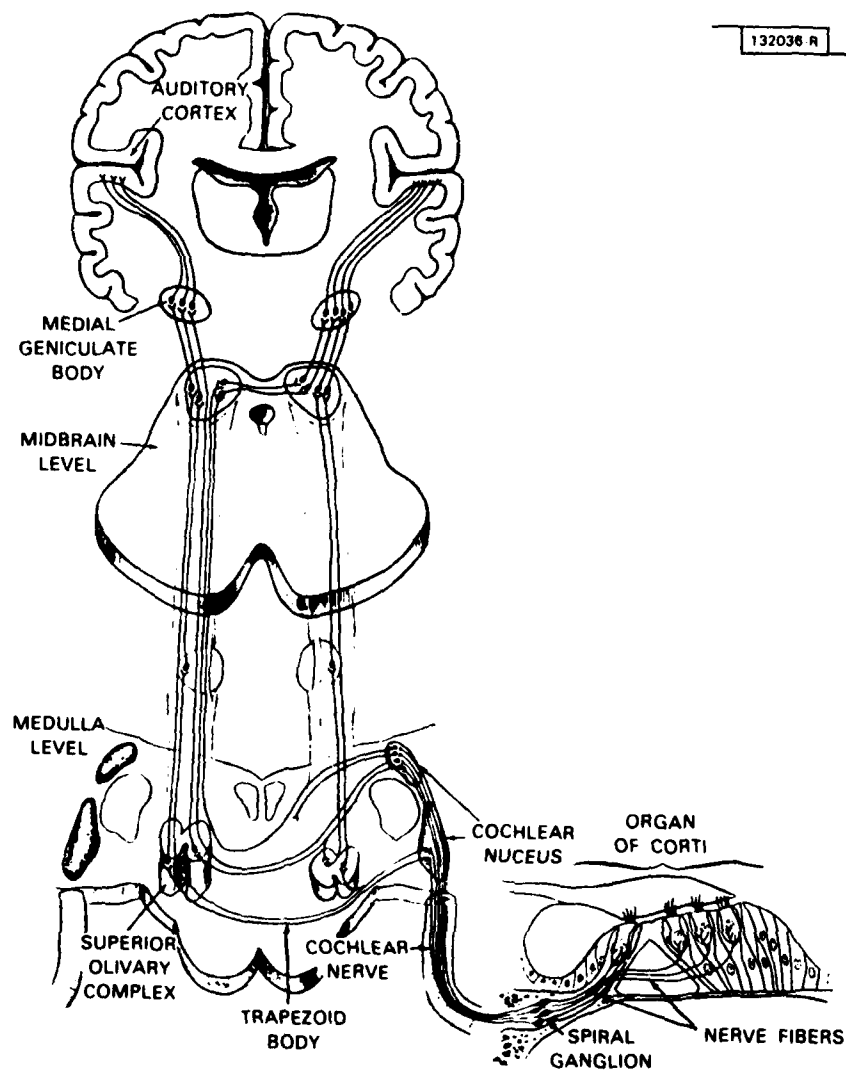


Fig. 17. Diagram of the auditory pathways linking the brain with the ear.

same ear. These pathways and crossovers occur at the level of the brain associated with the medulla. At the mid-brain level, nerve fibers travel from the superior olive and cochlear nucleus to a nucleus in the mid-brain called the inferior colliculus located at the top portion of the mid-brain. Again there are crossing fibers between the two ears so that binaural processing can occur. Finally, the next higher level of the auditory pathway moves to the medial geniculate body and then on to the auditory portions of the cerebral cortex. Once past the inferior colliculus there are no crossing fibers, so any processing associated with binaural inputs probably takes place below the level of the thalamus which is the brain area containing the medial geniculate body. That more and more complicated processing or association operations takes place as the signals move up toward the cerebral cortex is born out by Table V shown of numbers of neurons involved for each level of processing for the monkey. From the 30,000 nerve fibers, and therefore nerves as well involved with the cochlear nerve bundle of each ear, the signals end at the cerebral cortex where approximately ten million nerves are involved in final processing. At the level of the cortex, in spite of the amount of processing and crossing over that has taken place, there is a one-to-one relationship, tonotopic arrangement, between place in the cortex and frequency of acoustic input. This is true at least in the sense that if a place in the auditory cerebral cortex is excited by a microprobe into the exposed brain, the subject will hear a tone of corresponding frequency.

For completeness of this rather sketchy discussion, it is worth noting that pathways exist in the downward direction from cortex back to the hair

TABLE V
CELLS IN THE AUDITORY NUCLEI OF THE MONKEY*

Central Auditory Nucleus	Number of Cells
cochlear nuclei	88,000
superior-olivary complex	34,000
nuclei of lateral lemniscus	38,000
inferior colliculus	392,000
medial geniculate body (pars principalis)	364,000
auditory cortex	10,000,000

*Foundation of Modern Auditory Theory II - Tobias
Academic Press 1972. Chapter VI, p. 252.

cells. These pathways appear to be involved with AGC function as described earlier, but may have more sophisticated albeit unknown properties. At any rate, there are roughly 600 fibers in the cochlear nerve that terminate on the hair cells and provide this backward path to each ear from the level of the cortex.

Concluding Remarks

In this section, we have tried to outline the signal processing implemented by the organic structures known as the peripheral auditory system and the brain organized to deal with acoustic stimuli. We have pointed out that the basilar membrane implements a highly overlapped filter bank frequency analysis and the nerve fiber connections to the hair cell transducers provide information about this filter bank analysis in a very redundant manner to the brain. We also pointed out that the peripheral system provides high resolution in frequency at lower frequencies, and high resolution in time at the higher frequencies associated with acoustic inputs to the ear.

At the present time, our knowledge of the auditory system in terms of overall simulations useful in design of communications systems is sparse. However, even our limited knowledge suggests some properties vocoders and other speech equipment should retain or try to copy to provide robust and truly acceptable behavior from voice output to perception. These can be summarized as follows:

- (a) A large number of highly overlapped bandpass filters.
- (b) Many detectors at each frequency.
- (c) Final decisions based on probabilities.
- (d) Adaptivity to changing conditions.

VI. MODELLING THE AUDITORY SYSTEM

The ultimate goal of the perception-based vocoder program is to provide vocoder systems of improved intelligibility, quality and robustness. In this report, the robustness issue has centered about the performance of vocoders in noisy acoustic environments such as airplanes. Of particular interest to us has been the understanding of the behavior of the human auditory system. This interest is based on our major assumption that an analyzer designed to emulate the auditory system should result in a system with the desirable properties of high intelligibility, quality and robustness. Thus, one of our aims is to build software and/or hardware that emulates the auditory system.

The chain of sound leading to a percept begins at the entrance to the outer ear, travels through the middle ear, through the cochlea with its basilar membrane and hair cells, out to the primary auditory nerves, and through various way stations (cochlear nucleus, superior olive, lateral geniculate) on its way to the auditory cortex. Meaningful and sufficiently abundant physiological data are available only as far as the spike trains from the primary auditory nerve. Successful measurements from this nerve have been documented for various stimuli such as tone bursts, clicks and various synthetic speech-like sounds. Based on such measurements, various models for this part of the auditory system have been proposed.

A different approach to modelling is based on measured behavior. For example, numerous psychophysics tests on human pitch perception lead to models of the entire human auditory system but with the restriction that

this model need only emulate behavior as regards pitch perception. Other models are based on measured perception of various (usually synthetic) speech sounds.

Our first job will be to review some of these models. It is important to point out that all existing models are computational models, that is, computer algorithms that are (hopefully) computationally equivalent to measured behavior. In particular, although we know that the chain of anatomical connections beyond the cochlea consists exclusively of neural connections, it is the presumed end result of neural behavior rather than the neural behavior itself that is modelled. In this section, we propose the longer range approach, that models of auditory behavior (such as pitch perception) be modelled using neuron-like elements and that simulation programs of neural networks be incorporated directly into our models. Our hope is that as physiological measurements are made deeper into the auditory system, modelling along these suggested lines can remain more self-consistent and adaptable to new results.

Von Békésy [18] measured the response of the basilar membrane to various stimuli and thus set the stage for the filter bank approach to auditory models. Helmholtz [19] had already proposed that the ear had properties akin to that of a Fourier analyzer and Von Békésy showed that different places on the basilar membrane vibrated at different resonant modes. Based on such data, Flanagan [20] was able to extensively simulate the properties of the auditory response through the basilar membrane motion.

Kiang [21] and others have measured the output of the auditory nerve to various sounds. We know that the spike train produced by a single auditory nerve is to a great extent controlled by hair cell functioning, but it proves to be very difficult to perform good measurements on the hair cells of cats; Weiss [21a] and others have worked with the alligator lizard. Various models of hair cell-neuron transduction have been proposed [25].

Observations of spike trains from the primary auditory nerve (in particular, for the anesthetized cat) strongly suggest a statistical formulation to describe patterns of nerve firings. Siebert's [23] work has been most influential in this regard. In the absence of an auditory stimulus, nerve firings are random. When sound is impressed on the ear, the firings are influenced by the signal; a statistical description is still called for. Siebert invokes a non-stationary Poisson process with intensity function $r(t)$ and states that this statistical model fits much observed data except in one respect; "actual nerve records show, relative to a Poisson process, a deficiency of short interevent intervals. This phenomenon is usually attributed to a 'refractory' effect in the nerve fibers and/or receptors that prevents one firing from immediately following another."

Weiss' model of auditory nerve firings [22] (shown in Fig. 18) includes the refractory effect. In this model, whenever the membrane potential exceeds the threshold potential, the neuron will fire and will instantly raise the threshold potential to the high value R_M . This threshold will now decay exponentially with time constant T_R , so that

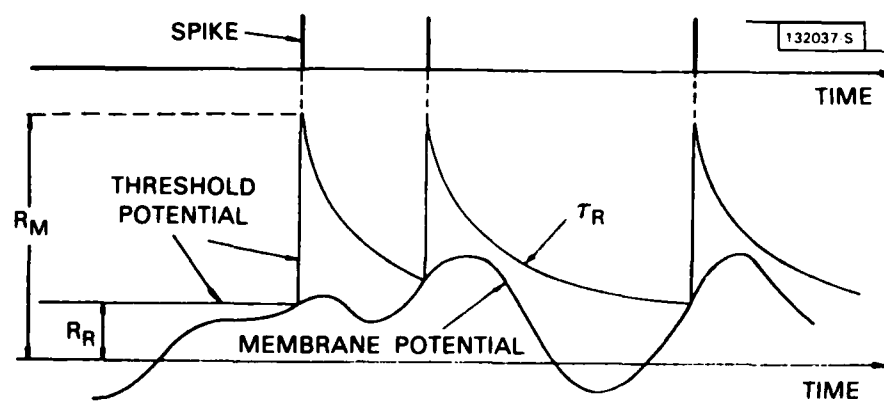


Fig. 18. Diagrammatic representation of membrane potential, threshold potential, and spike activity of the model neuron. From 59 Weiss.

R_M =maximum threshold potential

R_R =resting threshold potential

T_R =time constant of the exponential decay
of the threshold from its maximum to its
resting value

short intervals between adjacent firings have low probability of occurrence.

Other models of stimulus-induced activity of auditory nerve fibers have been proposed by Rose and Johnson [24,26]. A recently proposed model by Lyon [27] contains an interesting variation on the filter bank that is common to all such models. Lyon's model is shown in Fig. 19. The outputs are indicated by the small letters a, b, c, etc. A signal, in addition to being filtered, arrives at different outputs with different delays; in this way, the basilar membrane propagation velocity is taken into account. Lyon, by incorporating automatic gain control inputs from neighboring channels, also attempts to include lateral inhibition in his model. Lateral inhibition is an effective sharpening of sensory channel responses by weighting outputs with adjacent or lateral channels.

The remainder of this section will be devoted to a discussion of some preliminary work toward our own model of the auditory system.² Our goal is somewhat grandiose; we would like to develop a model that not only agrees with measured physiological data, but also can help explain many of the phenomena centered about pitch perception and spectral perception. Although this model is in a sense computational (since all simulation will be done by computer), we also intend the components of the model to consist of neuron-like elements once the stimulus gets to the auditory nerve.

The present simulation is being developed on the LDSP (Lincoln Digital Signal Processor) facility. Our first effort is directed toward pitch perception; for this reason, we have limited the bandwidth to Ritsma's "regional dominance" which lies approximately in the region 200-1500 Hz. Our 40 filters are each of bandwidth 140 Hz, and the center frequencies are

²The ideas presented here are in large part due to T. Bially.

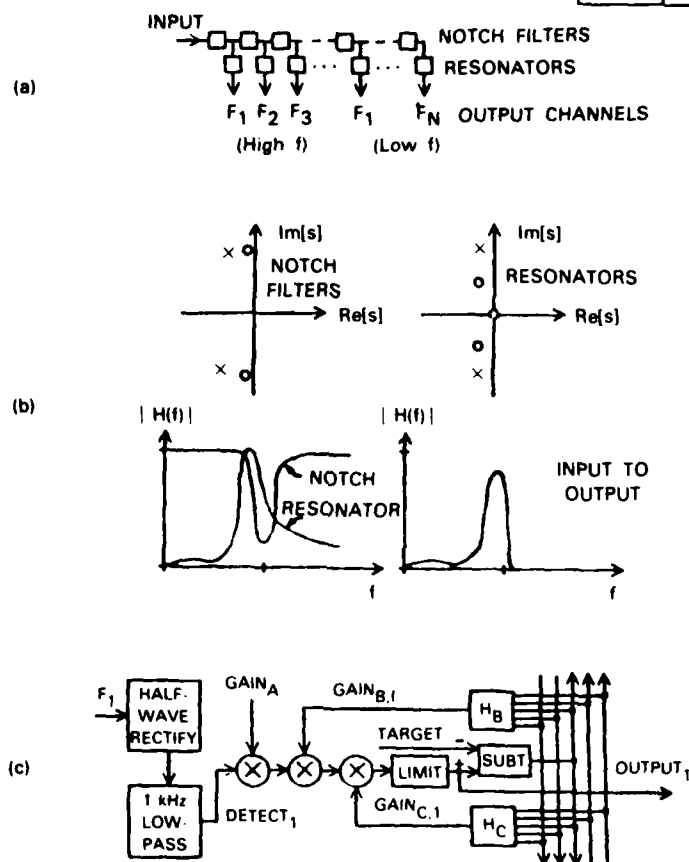
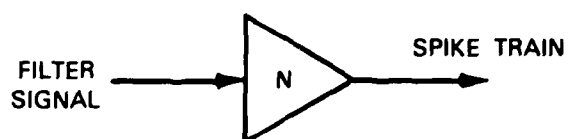


Fig. 19. (a) Block diagram of the cascade/parallel filterbank. (b) Pole-zero plots and transfer functions of filters used in the filterbanks. (c) Block diagram of one channel of the detection and compression models.

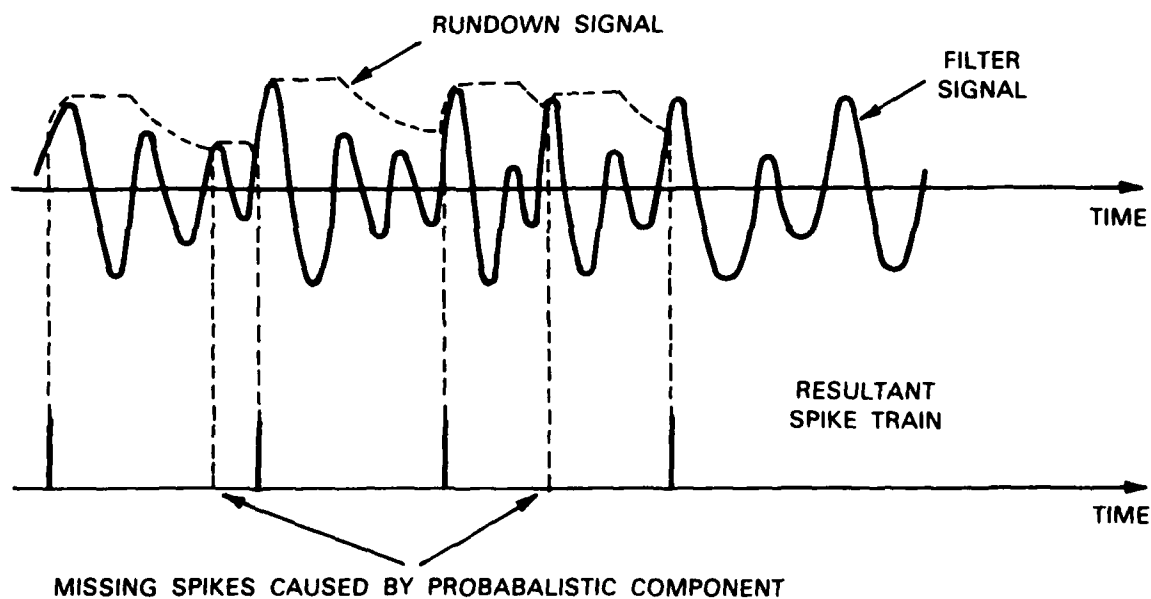
30 Hz apart. (In the future, we expect to bring our filter design more in line with measured data, perhaps following Lyon's filter model. Our present feeling is that the relatively large overlap among adjacent filters is the key ingredient.)

Many model makers have tried to incorporate hair cell activity directly into the model. We have taken a somewhat unorthodox approach and go directly from the bandpass filter outputs to the spike trains. In so doing, we have tried to incorporate many of the known features of auditory channels such as refractory period, adaptation, synchrony, and randomness. It is important to note that the spike trains obtained by stimulating a "neuron" with a filter output constitute only the first layer of nerve fiber signals. Our intention is to eventually proceed to a second and perhaps even a third layer so that the higher order processes of pitch and spectral perception can be modelled. This approach is more akin to some of the classical efforts to simulate neural networks that were carried out in the late 1950s by Clark and Farley [28].

Our model bears a resemblance to any of the six pitch detectors in the Gold algorithm [6]. A spike occurs when the rundown signal intersects the filter signal, as shown in Fig. 20. Following this, the rundown signal tracks the filter signal until the latter reaches a positive peak, at which time a refractory interval of constant duration is initiated. During this interval, no new spikes can be generated. Also during this interval, the random signal continues to track the filter signal so that it will move to a higher peak, should one occur. Once the refractory interval ends, the rundown signal returns to its normal mode of exponential decay, until it



(a)



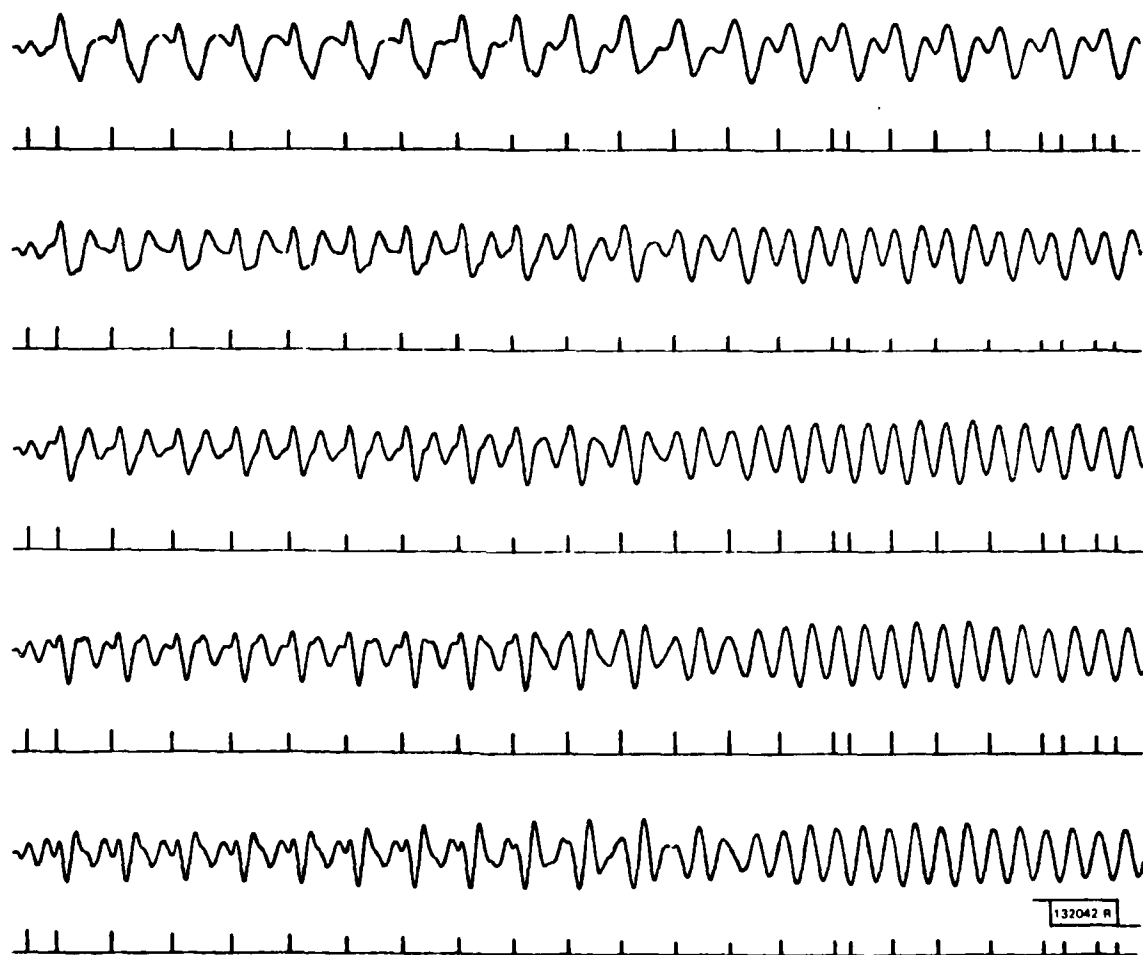
(b)

Fig. 20. Neuron-like elements in pitch measurement model.
 (a) Nomenclature for simple "neuron" N. (b) Operation of neuron N.
 (c) Level B of neural net.

again intersects the filter signal. When this occurs, a spike will be generated with a probability proportional to the ratio of the intensity (at intersection) to the intensity at the previous intersection. Introduction of this probabilistic component usually causes intervals between spikes to be multiples of the basic pitch period; this characteristic is reminiscent of interval histograms, as described in Section V.

The display of Fig. 21 shows some early results from this program. The top line shows 150 milliseconds of a speech waveform. The other signals show the outputs of highly overlapped band pass filters. Below each filter output is the spike train obtained by processing the signal with the neuron-like element described by Fig. 20. A cursory look at this data suggests that the spike trains from many neurons might be combined together to yield a cleaner and more reliable pitch sequence than is obtained from a single element. This step could presumably be done through the use of additional layers of neuron-like elements.

The program is being implemented on the LDSP. Because of the many processing elements involved, real time operation is not possible. However, we anticipate that by carefully analyzing the behavior of these elements in response to actual speech signals, we will begin to develop some insight into how their outputs might be combined in order to yield robust and reliable estimates of pitch and spectrum. We will then begin to consider additional layers of neuron-like elements to implement these processes.



21. Outputs of several filters and neurons for male utterance of length 150 msec. (See text for full explanation.)

REFERENCES

- [1] Speech Enhancement edited by J. S. Lim, (Prentice-Hall, 1983).
- [2] E. C. Cherry, "Some Experiments on the Recognition of Speech with One and with Two Ears," J. Acoust. Soc. Am., 25, pp. 975-979 (1953).
- [3] E. Singer, "A Comparative Study of Narrowband Vocoder Algorithms in Air Force Operational Environments Using the Diagnostic Rhyme Test," Technical Report 590, Lincoln Laboratory, M.I.T. (6 January 1982).
- [4] C. P. Smith, "Digital Voice Processor Consortium Interim Report," (August 1982).
- [5] G. A. Miller and P. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am., 27, pp. 338-352 (1955).
- [6] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Period of Speech in the Time Domain," J. Acoust. Soc., 63, pp. 498-510.
- [7] H. Fletcher, "Speech and Hearing in Communication," (Van Nostrand, New York, 1953).
- [8] J. F. Schouten, "The Residue, a New Component in Subjective Sound Analysis," Proc. Kon. Acad. Wetensch (Neth.), 43, pp. 356-365 (1940).
- [9] J. C. R. Licklider, "'Periodicity' Pitch and 'Place' Pitch," J. Acoust. Soc. Am., 26, p. 945 (A) (1954).
- [10] A. Seebeck, "Beobachtungen uber einige Bedingungen der Entstehung von Tonen," Ann. Phys. Chem., 53, pp. 417-436, (1841).
- [11] A. J. M. Houtsma and J. L. Goldstein, "The Central Origin of the Pitch of Complex Tones: Evidence from Musical Interval Recognition," J. Acoust. Soc. Am., 51, pp. 520-529, (1972).
- [12] J. L. Goldstein, "An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones," J. Acoust. Soc. Am., 54, pp. 1496-1516, (1973).
- [13] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," J. Acoust. Soc. Am., 71, pp. 1568-1580, (June 1982).
- [14] E. de Boer, "On the 'Residue' and Auditory Pitch Perception," Chapter 13 from Handbook of Sensory Physiology, 5, edited by W. Keidel.

- [15] G. A. Miller, and W. G. Taylor, "The Perception of Repeated Bursts of Noise," J. Acoust. Soc. Am., 20, pp. 171-180, (1948).
- [16] J. W. Hall III and R. W. Peters, "Pitch for Nonsimultaneous Successive Harmonics in Quiet and Noise," J. Acoust. Soc. Am., 69 (2), pp. 509-513, (Feb. 1981).
- [17] R. J. Ritsma, "Frequencies Dominant in the Perception of Pitch of Complex Sounds," J. Acoust. Soc. Am., 42, pp. 191-198, (1967).
- [18] G. von Békésy, Experiments in Hearing, (McGraw-Hill Co.), (1960).
- [19] H. von Helmholtz, On the Sensations of Tone as a Physiological Basis for the Theory of Music, (Dover), (1954).
- [20] J. L. Flanagan, "Models for Approximating Basilar Membrane Displacement - Part II," B.S.T.J., 41, pp. 959-1009, (1962).
- [21] N. Y.-S. Kiang, T. Watanabe, E. D. Thomas, and L. F. Clark, Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, (MIT Press, Cambridge, MA), (1965).
- [22] T. J. Goblick Jr. and R. R. Pfeiffer, "Signal Processing Characteristics of the Peripheral Auditory System," Technical Note 1966-50, Lincoln Laboratory, M.I.T., (1966), DDC AD645781.
- [23] W. M. Siebert, "Frequency Discrimination in the Auditory System: Place or Periodicity Mechanism," Proc. IEEE, 58, No. 5, (May 1970).
- [24] J. E. Rose, "Discharges of Single Fibers in the Mammalian Auditory Nerve" in Frequency Analysis and Periodicity Detection in Hearing, edited by R. Plomp and G. F. Smoorenburg, A. W. Sijthoff, Leiden (1970).
- [25] M. R. Schroeder and J. L. Hall, "Model for Mechanical to Neural Transduction in the Auditory Receptor," J. Acoust. Soc. Am., 55, pp. 1055-1060, (1974).
- [26] D. Johnson, "The Response of Single Auditory-Nerve Fibers in the Cat to Single Tones: Synchrony and Average Discharge Rate," Ph.D. Thesis, M.I.T., (1974).
- [27] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," ICASSP '82, pp. 1282-1285.
- [28] W. A. Clark and B. G. Farley, "Generalizations of Pattern-Recognition in a Self-Organizing System," Proc. 1955 Western Joint Computer Conf.

- [29] B. Delgutte, "Representation of Speech-Like Sounds in the Discharge Patterns of Auditory Nerve Fibers," J. Acoust. Soc. Am., 68, pp. 843-85 (1980).
- [30] M. B. Sachs and E. D. Young, "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve," J. Acoust. Soc. Am., 68, pp. 858-876, (1980).

APPENDIX A

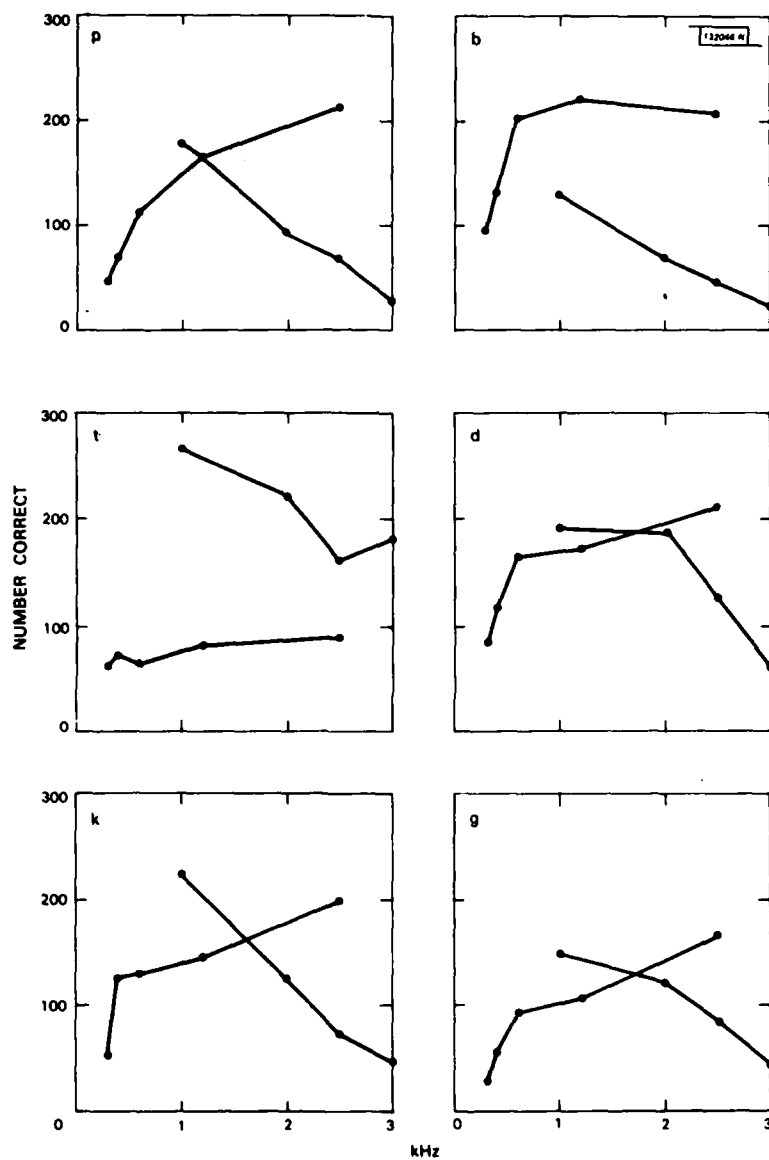


Fig. A-1. Number of correct plosive identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).

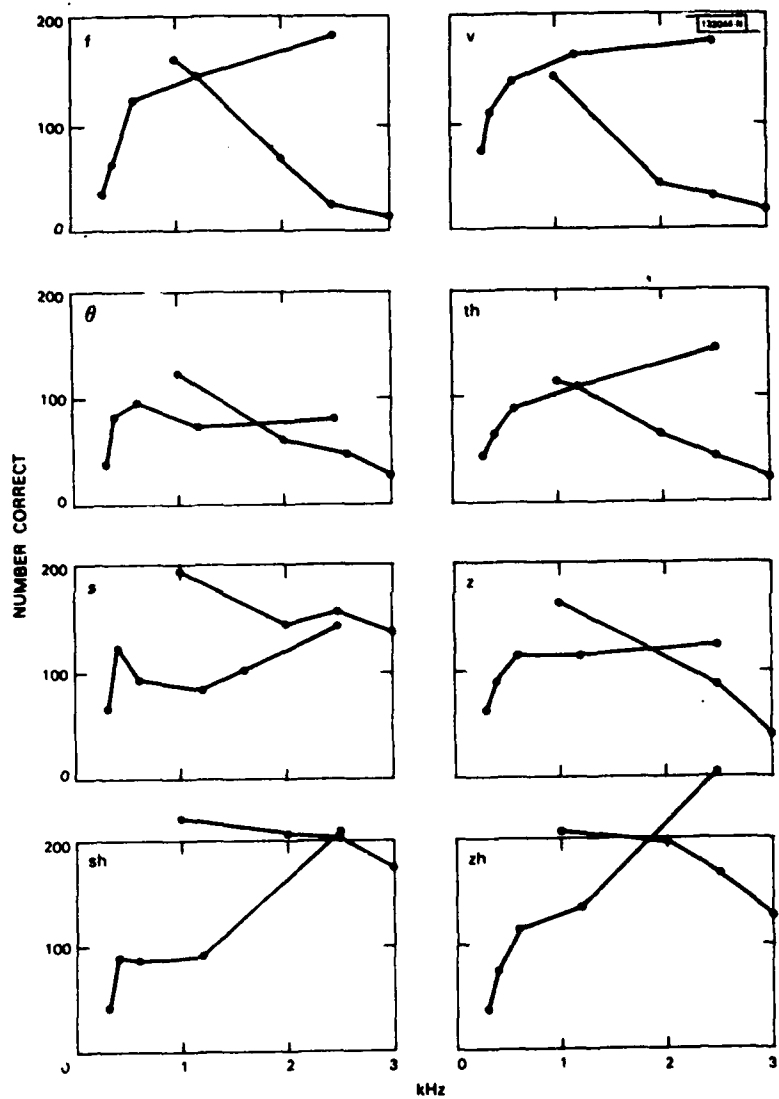


Fig. A-2. Number of correct fricative identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).

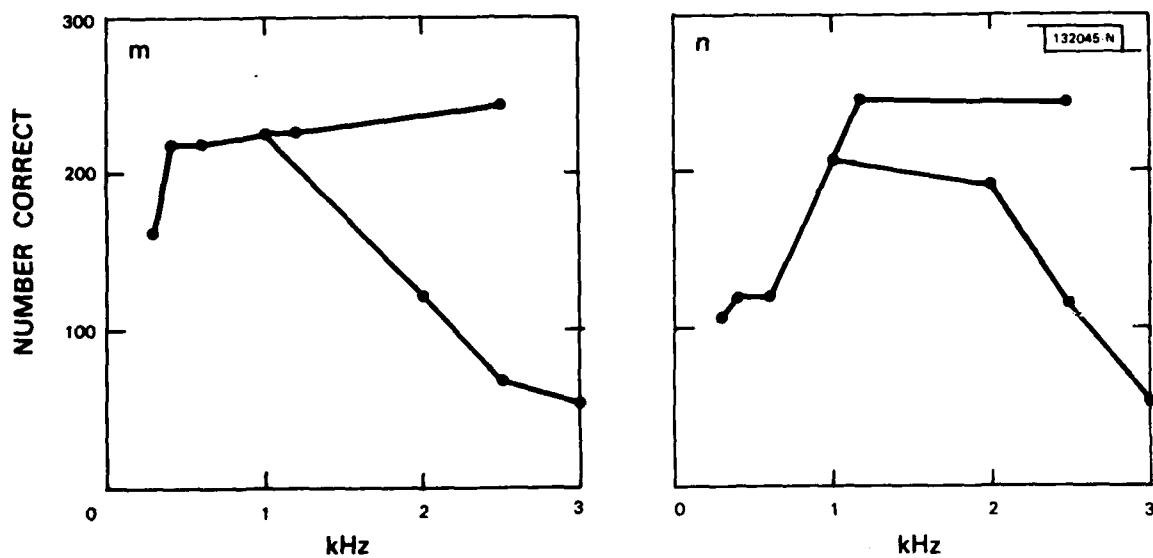


Fig. A-3. Number of correct nasal identifications vs. low pass and high pass cut-off frequencies (from Miller-Nicely).

APPENDIX B

Confusion Matrix for 5 KHz RAW (Item 2 of Table III)

77

Confusion Matrix of flatvoc (Item 3 of Table III)

78

Confusion Matrix of flatvoc (Item 20 of Table III)

	p	t k	b	d g	f	θ s sh	v	th z zh	w r y l	m n	ch j	h
P		9 10	1		5							
T	33	10	6		12							
K	15	11	6								10	
B	7		15				34			3		
D	5		11	6			13				29	
G			4	3								
F	13						9					8
Θ					42							
S					17			7				
SH												
V			11		13							
TH			16					1				
Z												
ZH												
W											4	
R									1	2		
Y												
L												
M			3							12		
N				2						3		
CH		3				1					3	
J				13								
H					1							

Confusion Matrix of M2400 (Item 5 of Table III)

80

Confusion Matrix of M2400 (Item 12 of Table III)

[illegible]

APPENDIX C

APPENDIX C
 DESCRIPTIONS CATEGORIZING DRT INTELLIGIBILITY
 (from Reference 4)

DRT Intelligibility Category Score	Examples (based on three male speakers)	
100	Excellent	Unfiltered speech In quiet environment; no 4 kHz low-pass significant distortion or filtered speech reverberation; high-quality microphone
96	Very good	CVSD at 32 kbps: Ber<1% speech in quiet environment
91		CVDS at 16 kbps: Ber<1%
87	Good	APC at 9600 BPS Speech from CONUS median analog voice channel quiet LPC-10 narrowband vocoder at 2400 BPS Environment
83	Moderate	LPC-10 narrowband vocoder at 2400 BPS Speech in jet A/C BER <1% Cabin noise (87 db SPL)
79	Fair	LPC-10 at 2400 BPS: 2% BER Speech in quiet environment (no error correction)
75	Poor	LPC-10 at 2400 BPS: 5% BER Speech in quiet environment (with bit error correction)
70	Very poor	Experimental 800 BPS voice processor: speech in quiet (zero bit errors)
	Unacceptable	LPC-10 at 2400 BPS: Speech from helicopter noise environment

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-83-226	2. GOVT ACCESSION NO. H138660	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Vocoder Analysis Based on Properties of the Human Auditory System		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER Technical Report 670
7. AUTHOR(s) Bernard Gold and Joseph Tierney		8. CONTRACT OR GRANT NUMBER(s) F19628-80-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173-0073		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element No. 33401F Project No. 7820
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331		12. REPORT DATE 22 December 1983
		13. NUMBER OF PAGES 96
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Electronic Systems Division Hanscom AFB, MA 01731		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES None		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) vocoders auditory modelling diagnostic rhyme tests perception psychophysics		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) When a person listens to speech corrupted by noise or other adverse environmental factors, speech intelligibility may be impaired slightly or not at all. The same corrupted speech, after being vocoded, often causes drastic intelligibility loss. This is due to the fact that the human peripheral auditory system is a superior signal processor to that of the vocoder. This report is based on the premise that a vocoder analyzer that better resembles the peripheral auditory system would function in a superior manner to present-day vocoders. Topics include reviews of speech enhancement techniques, perceptual analysis of diagnostic rhyme test data, a brief description of the peripheral auditory system and an outline of proposed psychophysical tests. The final section is devoted to a discussion of some preliminary work on computer simulation of an auditory model.		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)