Monograph:
Kirby, H.R. and Roach, P.J. (1987) Voice Degradation in Using Speech Recognisers for Transcribing Inventory Data: Draft Final Report. Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 236

Published paper

Kirby, H.R., Roach, P.J. (1987) *Voice Degradation in Using Speech Recognisers for Transcribing Inventory Data: Draft Final Report.* Institute of Transport Studies, University of Leeds. Working Paper 236

Working Paper 236

January 1987

VOICE DEGRADATION IN USING SPEECH RECOGNISERS FOR TRANSPCRIBING INVENTORY DATA: DRAFT FINAL REPORT

H R Kirby

and

P J Roach

1. FRONT COVER

AD

VOICE DEGRADATION IN USING SPEECH RECOGNISERS FOR TRANSCRIBING INVENTORY DATA

DRAFT VERSION OF FINAL TECHNICAL REPORT

by

H.R. Kirby and P.J. Roach

January, 1987

United States Army

EUROPEAN RESEARCH OFFICE OF THE U.S. ARMY

London England

CONTRACT NUMBER DAJA45-85-C-0043

The University of Leeds

Leeds LS2 9JT, U.K.

2. TITLE PAGE

(WILL BE SUPPLIED BY U.S. ARMY)

3. ABSTRACT

It is known that problems arise in long sessions of voice tape-recording for off-line data entry to computers via speech recognition systems, as a result of operator fatigue or loss of attention. In this study the task of reading vehicle licence plates aloud for one hour was simulated in laboratory conditions, each speaker undergoing one recording session with feedback on recognition accuracy and one without feedback. No significant difference in recognition success rates between the two conditions was discovered. The audio tape-recordings of the sessions were analysed acoustically for fatigue-induced changes both in long-term prosodic characteristics (including fundamental frequency, intensity, spectral balance and rate) and in segmental characteristics such as frequency of occurrence of different sound types and segmental durations. Although a number of intra-speaker differences were detected in various measures, no consistent tendencies were found in all speakers. It is concluded that the choice of speakers and conditions resulted in insufficient fatigue to produce clear-cut effects; however, it is felt that the more sophisticated techniques developed during the project on the basis of the automatic segmentation of continuous speech is capable of more revealing analysis than the relatively crude techniques used previously. It is not felt likely that this would result in automatic techniques for improving speech recognition accuracy, but it could form the basis for more effective operator training and assessment of new applications and techniques.

4. KEYWORDS

SPEAKER RECOGNITION

VOICE DEGRADATION

TRANSPORT STUDIES

SPEECH INPUT

SPEECH RECOGNITION

PHONETICS

HUMAN FACTORS

TRAFFIC SURVEYS

5. CONTENTS

6. LIST OF ILLUSTRATIONS AND TABLES

SECTION 7 : PROJECT DESCRIPTION AND FINDINGS

1 INTRODUCTION

The technology for speech input to computers has reached a level of development which makes it fully practical for use in a wide range of real-life applications in which the user's eyes or hands are not free for normal writing or keyboarding; certain kinds of moving-traffic surveys constitute tasks of this kind. The principal disadvantage of speech input has until very recently been the high cost of reliable equipment, and this problem has been found particularly acute because in most applications each human operator would have to have exclusive use of an entire speech recognition system. However, one way of making more economical use of speech input technology is to have a number of operators working with a single recognition system by recording data "off-line" on audio tape and later playing that data through the recogniser for computer input. The object of the present study has been to look at some of the problems arising from this way of using speech input technology, and to make recommendations for maximising its efficiency. As far as is known, work on the specific problems associated with recognition of taped speech input is only being done at Leeds University and at the Construction and Engineering Research Laboratory at Urbana-Champaign, Illinois.

Previous work in the Institute for Transport Studies (I.T.S.) at Leeds University was done under the direction of H.R.Kirby and P.W.Bonsall, funded by the Science and Engineering Research Council (Kirby, 1983[1]; Bonsall, Hardwick and Kirby, 1985[2]; Kirby and Bonsall, 1985[3]). The S.E.R.C. grant enabled two different types of automatic speech recogniser to be assessed for this purpose. One kind could recognise isolated words, while the other could cope with phrases. The assessment included both laboratory and field trials of the recognisers and their associated equipment (microphones and tape recorders). Different kinds of training programme were evaluated and comparisons were also made with other methods of data collection and transcription. The main application evaluated was that of registration number surveys, which are typically used to estimate journey times along a street or through a network, or to indicate the route used. One of the S.E.R.C. project's findings was that aspects of the speaker's voice change during recording sessions, apparently causing a loss of recognition accuracy. In the traffic-related conditions studied by I.T.S., the problem was thought to be due to changes in volume and pitch induced by stress under varying levels of traffic noise and flow; in the inventory recording conditions studied by Urbana-Champaign, the problem was thought to be due to hoarseness or fatigue.

The measurement and assessment of long-term changes in the voice of a speaker present many problems for the speech scientist. One of the main objectives of this study has been to develop simple and robust measures capable of detecting such differences in normal (non-pathological) voices.

Two major problem areas have been explored:

(i) In long spells of work all speakers eventually exhibit signs of fatigue: in what way(s) are the signs of vocal fatigue physically manifested in the speech signal, how is the efficiency of speech recognition by machine impaired and what can be done to minimise the impairment?

(ii) All speakers are different: are some speakers more effective in using speech input than others?

We have throughout this study of variability in speaker performance attempted to keep conceptually separate three different factors: human, machine and environmental. This is by no means a simple task: some of the problems are outlined by Martin & Welch (1980)[4] and Lea (1980)[5]. Human factors which may affect the performance of the operator include such things as acceptance of the microphone mounting with regard to comfort and appearance and also confidence in the system which is being used. Wilpon and Roberts (1986)[6] report that experienced speakers are more consistent than naive users, and that male speakers have better performance than female speakers. Machine factors such as bandwidth limitations of the transducer may not only have a direct effect on the input, but also increase the task complexity leading to more rapid deterioration in the performance of the operator. There may also be the added complication of fluctuations in environmental noise level which could corrupt the input to the speech recogniser. There are various techniques for speech enhancement and bandwidth compression of speech degraded by background noise, and a discussion of these is given by Lim and Oppenheim (1979)[7]. However, investigation of this field is outside the scope of the present project.

2 DATA FOR THE STUDY

It was initially planned to use three sets of recorded data:

(i) Recordings of genuine inventory data made under "real life" conditions by U.S.Army personnel or U.S.Army-funded research staff.

(ii) Existing recordings of traffic surveys made by the Institute for Transport Studies.

(iii) Simulated data recorded under laboratory conditons for controlled experimentation.

In practice, we were unable to use the data we wanted: the data in (i) above could not be made available to us, while (ii) were felt by the Linguistics & Phonetics researchers not to be suitable as a database for detailed acoustic examination. This

was because they had been made in very adverse recording conditions and, since at the time no laboratory study of the type reported here was envisaged, some of the relevant factors were not controlled for: the recordings are of different lengths, and there is insufficient information about the speakers and their success rates. It was therefore decided that the only way to carry out controlled experimental investigations within the limited time of the project would be to record simulated data under laboratory conditions ((iii) above). A number of exploratory test sessions were carried out to devise a suitable experimental technique.

It was decided to simulate the reading of vehicle number plates, as in a moving-traffic survey, and to require subjects to continue the task without a break for a one-hour period. This was felt likely to be long enough to produce signs of fatigue; since all subjects were required to do two recording sessions, it was felt unlikely that many subjects would tolerate longer sessions than one hour. A computer-generated screen display of random vehicle types and number plates was produced with a microcomputer, and the program allowed random intervals between presentations of data items, with the option of simulating different traffic flow rates (i.e. number of vehicles per hour). The speech was simultaneously tape-recorded. The task for the subject was to read the items put on the screen, which were made up of the following:

(i) vehicle type (car, van, lorry, motorcycle)

(ii) up to three numbers (each to be read as a separate item)

(ii) one letter, to be read out according to international conventions (A="alpha", B="bravo", etc.).

The rate chosen was 600 vehicles per hour, so that each 1-hour recording contains approximately 600 simulated number-plate readings. We had originally planned to make some recordings in which speakers were subjected to different amounts of environmental noise during their reading tasks, but this was found impossible during the short time-scale of the project.

Each subject made two recordings, under two different conditions: in one, (s)he could see the recogniser read-out during the recording, and thus had immediate feedback on any errors made, while in the other the machine read-out was obscured and the subject received no feedback. It is interesting to compare the study by Wilpon and Roberts (op cit; see also Roberts et al, 1986[8]) in which the feedback was instead given by a graphical representation ("barometer-style") of the distance between the current input word and the selected template.

A total of five subjects (three female and two male) were recorded in this way. No instructions were given to subjects about speaking style other than the advice that they should speak naturally. This was because several studies (e.g. Bobrow and Klatt, 1968[9]) have indicated that speech is more consistent if

subjects are not instructed to adopt a particular manner of speaking. The speakers were also given trial runs with the equipment before the actual recording was made.

3 EQUIPMENT AND FACILITIES

Two types of word recogniser were available to the project: the Interstate Electronics Corporation SYS300, distributed through KODE, and the Marconi SR-128X manufactured by Marconi Space and Defence Systems Ltd. The former is only capable of recognising words spoken in isolation, while the latter can recognise words in short phrases spoken without internal breaks. It was clear from the outset that of the equipment available to us, the Marconi SR-128X recogniser was by far the more reliable and accurate, and the other system was not used.

Audio tape recordings were made on cassette with a professional-quality Sony Walkman recorder. The complete set of recordings is submitted with this report.

The acoustic analysis was carried out in the Phonetics Laboratory of the Department of Linguistics & Phonetics. For the long-term measurement of fundamental frequency we used a Frokjaer-Jensen Fundamental Frequency Meter, type FFM650, and for long-term sound pressure level a Frokjaer-Jensen Intensity Meter, type IM360. For broad-band spectral analysis we used equipment specially constructed for the purpose in the Phonetics Laboratory Workshop: this was designed to receive a tape-recorded audio signal and to produce four voltage-varying outputs representing filtered, rectified and integrated amplitude in different frequency bands. The four "traces" were given filter settings that had been found useful in previous work (see Section 4.3.2 below). These instruments were connected to the analog input ports of a BBC Microcomputer for computer logging of the instrumental output. The effective maximum sampling rate of this device on a single channel is approximately 100 per sec, which is acceptable for gross and long-term measures but is inadequate for fine-grained acoustic analysis (for which a minimum rate is normally reckoned to be 10,000 per sec). Sampling of four channels took 40 msec.

For the final series of experiments we made use of a PDP-11/73 minicomputer system in the Linguistics & Phonetics Department. The software used comprised the ILS interactive signal processing package (Signal Technology, Inc.) and the LUPINS speech segmentation and labelling package developed in the Department. Further details of this are given below (Section 4.3.2). It is disappointing to have to report that work on this aspect of the project has been hampered by a long series of technical problems with the computer system, arising from incompatibilities between the I.L.S. software and the particular hardware configuration chosen; this resulted in inability to handle high-speed analog input-output until substantial modifications were carried out, and this was completed only at the end of the present project.

4 EXPERIMENTAL INVESTIGATIONS

4.1 Auditory Evaluation of the Data

When the primary database of ten hours of speech had been recorded it was felt that it would be valuable to have the opinions of some experienced listeners. An audience of 12 phonetically trained listeners was brought together and asked to listen to one-minute extracts taken from the ten recordings; two extracts were taken from each recording, one from near the beginning and one from near the end, giving a total of 20 items for the listening session. Although no formal test was carried out, since this would have detracted from the informal atmosphere of the listening session, it appeared that the audience, as well as the experimenters, could in many cases identify which passages came from the end of a recording and which from the beginning. The listeners gave written comments on their impressions of the recordings with specific reference to differences between beginning and ending samples: the most frequently oberved characteristics are listed below.

(i) Mean pitch appeared to have changed over time, though for some speakers in a generally downward direction and for others generally upward.

(ii) Intelligibility was felt to be lower in the ending samples, with weaker, less precise articulation.

(iii) There was a higher incidence of the voice quality known as "creaky voice" in ending samples.

(iv) Pitch range was narrowed towards the end.

(v) There were more abrupt changes in loudness at the end.

(vi) Low-falling pitch glides descended less far and less rapidly in ending samples.

These observations were kept in mind during the acoustic analyses. One crucial factor, however, was noticed by the experimenters, and that is that throughout the recordings the speakers appeared constantly to be "pulling themselves back" when their speech began to deviate from their non-fatigued norm, so that the speech degradation appeared to be cyclical rather than progressive. Presumably our recording sessions never reached the point of fatigue where our speakers stopped trying to be accurate altogether (see next section).

4.2 Success Rate

Each recording was checked for accuracy with regard to correct identification by the recogniser of each test item. Occasional reading errors by the human subjects were disregarded in this count, since what was under investigation was machine performance on error-free input data. Any test item (i.e. vehicle type and

number-plate data) whch contained any misidentification was counted as one error. For each recording, errors were counted over successive ten-minute periods and expressed as percentages of the total number of test items read during that period. The results are given in Table 1 and it is clear that, taking the group as a whole, no significant differences are to be found between with-feedback and without-feedback conditions, nor does any consistent trend to lower success rates over the one-hour recording period show up. It was therefore necessary to conclude that we had failed to produce a fatigue effect in our subjects powerful enough to cause a consistent drop in recognition accuracy. Presumably our speakers were too highly motivated, too comfortable and too interested in the experiment. It should not, of course, be concluded from the success rate scores that no changes took place in the speakers' voices during the recording sessions; however, if there were changes they were either too slight or too momentary to affect the machine's performance when averaged over a ten-minute period. The performance of a machine as sophisticated as the SR-128X would not, of course, be expected to show a linear falling-off of performance in relation to a progressive change in one or more speech parameters; it is consequently very difficult to produce a controlled degradation in speech recogniser performance.

TABLE 1: NORMALISED (%) ERROR RATES FOR SPEAKERS WITH AND WITHOUT VISUAL FEEDBACK

			MINS FROM START					
SPEAKER		0-10	10-20	20-30	30-40	40-50	50-60	mean
1.	with f/b 7.8	9.3	10	6.5	3.1	3.3	6.7	
	no f/b 5.5	5	8.3	4.4	10	8.3	6.9	
2.	with f/b	6.5	9.6	10	9.5	9.1	9.6	9.0
	no f/b 7.1	9.6	10	8.7	9.8	9.6	9.1	
3.	with f/b	7.7	7.7	5.9	10	5	6.3	7.1
	no f/b 3.5	6.4	7.8	7.1	10	2.1	6.1	
4.	with f/b	6.8	9	10	9.3	9.6	7.7	8.7
	no f/b 9.5	5.5	8.2	5.9	10	9.5	8.1	
5.	with f/b	6.2	6.4	7.6	8.2	8.2	10	7.8
	no f/b 4	8	8	10	8	8	7.7	

In Table 1, speakers 1 and 5 are trained phoneticians with 18 and 30 years, respectively, of professional work in speech. Speaker 3 was trained and experienced in the use of a speech recogniser, but had no phonetic training. Speakers 2 and 4 had a moderate amount of phonetic training but practically no experience of work with a speech recogniser. It can be seen that the average error rates of the "experienced" speakers are lower than those of the less experienced speakers, though we believe that with a sample of this size it would not be meaningful to employ tests of statistical significance. Speaker 1's performance does seem to have improved during the "with feedback" recording, while Speaker 5 seems to have done the opposite. Speaker 3's "without feedback" performance is the only one to exhibit something like the behaviour that we would have predicted, in that the error rate rises progressively up to the 50-minute mark; in the final ten minutes, however, the error rate drops sharply for reasons that we can only guess at. All the female error rates are higher than the male error rates, and this supports the finding of Wilpon and Roberts quoted in Section 1.4 above. We suspect that much of the development work on instrumentation used in speech research is based predominantly on male voices (for example, the choice of filter bandwidths in the original sound spectrograph is said to have been made on this basis), and this might result in a built-in tendency to favour male voices.

In anticipation of being provided later with authentic tape-recorded data from U.S. sources, we continued to work on the analysis of our own database to develop and test our research techniques; we also looked in the recordings for acoustic evidence of changes in speakers' voices (though the recogniser success rates suggested that any changes we found were likely to be small ones). At the conclusion of this work, we feel that it would have been worthwhile to record some trained speakers deliberately trying to simulate (and ultimately to exaggerate) various progressive fatigue-type changes in their speech on a speaking task similar to that used for our original recordings. This would have enabled us to provoke catastrophic failures in the machine's performance.

4.3 Instrumental Measures

Two types of instrumental measurement were used in the analysis of the recorded data: one was the long-term measurement of prosodic aspects of speech, and the other involved looking for changes over time in the segmental properties of speech. As will be seen, the conventional distinction between prosodic and segmental properties is not as clear-cut as is sometimes supposed.

4.3.1 Prosodic features: these are characteristics of speech which are constantly present and observable while speech is going on (Roach,1983[10]). The features we decided to examine were the following:

(i) Fundamental frequency (changes in mean value over time; changes in variance).

(ii) Gross spectral characteristics.

(iii) Mean sound pressure level.

(iv) Speed of utterance.

The instrumentation used is described in Section 3.2 above. The results given for measurements are mainly as reported in Dew et al (1986[11]), though the presentation and layout is somewhat different.

4.3.1.1 **Fundamental Frequency:** the first prosodic factor examined was long-term fundamental frequency (from now on, fundamental frequency is referred to by its standard abbreviation, F0). Despite the auditory impression of upward or downward movement in some recordings, no such movements were observed as long-term trends in the plots. However, it should be remembered that auditory impressions are based on a complex set of physical parameters and one should not expect to find anything like linear relationships between individual parameters and subjective attributes.

To produce a long-term record, the output of the F0 meter was sampled for successive periods of 4 sec, each yielding approximately 400 samples. Silences and spurious values below 50 Hz were ignored, and if the entire 4 sec stretch was found to have been silent (a rare event, though possible), a new 4-sec sampling period was started. Sampling then stopped and 2 sec was allowed for the computer to calculate the mean and standard deviation for the 4-sec "window" and to plot the results on the computer screen. The plot took the form of a vertical line representing 1 standard deviation above and one standard deviation below the mean. Informally, we can say that the plots in Figs. 1 through 5 represent mean F0 and range of F0 at successive points in time through a one-hour recording (the process was in fact allowed to run for a little over an hour, as can be seen from the time-base). At the end of the recording the screen contents were dumped to a dot-matrix

printer.

We now feel that a 4-sec window was too short, and may well have allowed too many spurious meter readings caused by speech onset and offset to enter the calculations. Other researchers (e.g. Steffan-Batog et al, 1970[12]) have used sample durations of around 1 minute; Hiller et al (1984[13]) used 5 seconds as an initial sample duration but then incremented the sample size in 5 sec steps up to 1 minute. However, all analog F0 extraction techniques produce spurious readings at points of uncertainty, and the only ways to avoid this effectively are either (i) to derive the measurement of vocal fold adduction and abduction electrically by means of a device (an "electroglottograph" or "laryngograph") which monitors trans-glottal impedance (Fourcin and Abberton, 1972[14]), or (ii) to use a computer system sophisticated enough to "know" what are plausible and what are implausible values in a given context. The former is unsuitable for our purposes, since all speakers would have to wear electrodes on their throat when recordings were taking place; the latter is now available in our laboratory, but has been so only since the experimental work on this project had to be brought to a close. The problems, and the computational means for overcoming them, are discussed in Leon and Martin (1972[15],) Johns-Lewis (1986[16]) and Hiller et al (op cit).

The plots all show gradual medium-term shifts of mean F0 and range, typically over a stretch of around 200 sec, as well as many abrupt discontinuities. However, none of the plots shows any long-term shift that lasts longer than 8-10 mins, unless one counts the gradual rises in mean F0 at the beginning of Speaker 2's "no-feedback" recording (Fig.2, lower) and at the end of Speaker 3's "no-feedback" recording (Fig.3, lower).

4.3.1.2 **Gross Spectral Characteristics:** the quality of a speaker's voice may change as a result of fatigue. It can happen that a difference in the mode of vibration of the vocal folds affects the overall spectral shape of the voice source waveform, and this could be reflected in changes in relative intensity in higher and lower regions of the spectrum. Vilkman and Manninen (1986[17]) found an increase in high-frequency energy in the speech of subjects working in several combinations of stressful conditions. It could also be possible that high-frequency energy produced during fricative sounds might decline through fatigue more noticeably than low-frequency energy produced during voiced sounds. It was thought unlikely that such differences would be observable in our very long-term plots, but it was felt to be worth looking at. The four-channel filter bank decribed in Section 3.2 was used, and the output data logged by microcomputer. Fig. 6 shows typical results for one male speaker (Speaker 3), but it was not felt that anything useful was learned from these plots. What would be far more meaningful would be to carry out long-term logging of spectral balance exclusively of sections of speech identified as one particular sound type, given that the sound is one of those that reaches something approaching a steady state (i.e. vowels, nasals and fricatives). The possibility of doing this in our laboratory now exists with the system described in Section 4.3.2 below, but it has not

been possible to do this within the time-scale of the present study.

4.3.1.3 Mean Sound Pressure Level: it has not yet proved possible to produce meaningful results for this parameter, though on the face of it this presents the simplest task. Plots of mean sound pressure level (henceforth s.p.l.) were produced and showed apparently random fluctuations rather similar to those of the F0 means described in Section 4.3.1.1 above. However, we suspected, on the basis of our experience in analysing intensity meter traces, that the data would be constantly varying over an extremely wide range, even on a dB scale, and when the "+/- 1 s.d." plotting technique used for the F0 plots in Figs. 1 - 5 was tried, the magnitude of the variance was such that little or nothing could be deduced from the plots. As in the case of the attempted analysis described in Section 4.3.1.2 above, we feel that a far more meaningful picture would emerge from long-term logging of selected portions of the speech on the basis of categorisation into different sound types. The intensity measures in the work of Vilkman and Manninen (op cit) illustrate this, being based on vowels extracted from read speech. Again we were prevented from doing this over long samples of speech by the delayed availability of our computerised analysis system. However, a small-scale "manual" analysis was carried out to see if the intended automatic system would be likely to show any differences. Only one speaker was used because of the extremely time-consuming nature of manual analysis. One minute was taken from near the beginning and one from near the end of each of two one-hour recordings, one made in the "with-feedback" condition and one in the "no-feedback" condition. The peak s.p.l. of every /s/ fricative found was measured. The results are given in Table 2.

TABLE 2: PEAK INTENSITY LEVEL (dB) OF /s/ H.P.FILTERED AT 3.9 kHz

	WITH FEEDBACK		WITHOUT FEEDBACK	
	near start	near end	near start	near end
MEAN:	19.05	17.04	16.21	14.28
S.D.:	1.59	2.33	2.09	1.92
N:	17	22	19	21

A Mann-Whitney U-test was carried out on the data. For both experimental conditions there was an overall reduction in sound pressure level between the beginning and the end of the recording, and this gave a significance level of 0.0081 for the first condition and 0.0113 for the second. If this proves to be a general tendency for other speakers and for other sound types it would be consistent with the observation that speech is less clearly articulated as time progresses.

Examination of individual words containing the /s/ phoneme does, however, suggest that in our data the difference in sound pressure level of this fricative has no direct bearing on the success or failure of recognition of the words in question.

4.3.1.4 Speed of Utterance: It would clearly be desirable to be able to have some measure of how rapidly a speaker was speaking at a given point in time, but this is something that is difficult to do. Traditionally, this has been done manually by making an oscillographic record of the speech to be measured and then counting the number of occurrences of some linguistic unit found in some unit of time. The coarsest measure used is normally words per minute; at the finer level of syllables per second it is usual to discount silences over a certain length, and the same is true of a count of phonemes or phonetic segments per second (Lehiste, 1979[18]). A second measure of speed of utterance is to measure the duration of some known linguistic unit at different points of time to see if it changes. Since our recordings contained only short bursts of speech, which often contained silences and hesitations, we felt that our speakers could not be considered to have settled to a particular speaking rate, and we therefore decided to confine ourselves to the second type of measure. However, given more time it would probably be worth experimenting with ways of measuring syllable and segment rates in data such as ours. In this section we describe a manually measured pilot experiment, and report on a study of speech segment duration in the next section.

The overall duration of selected words was measured manually from oscillographic traces. The words were taken from one-minute passages extracted from near the beginning and near the end of one-hour recordings. The results are shown in Tables 3 and 4:

TABLE 3: DURATION (ms) OF THE WORD "FOUR"

	WITH FEEDBACK		WITHOUT FEEDBACK	
	near start	near end	near start	near end
MEAN:	423	300	362	378
S.D.:	39.3	56.1	40.86	26.83
N:	6	5	5	5

TABLE 4: DURATION (ms) OF THE WORD "NINE"

	WITH FEEDBACK		WITHOUT FEEDBACK	
	near start	near end	near start	near end
MEAN:	401	284	338	324
S.D.:	22.5	40.8	32.5	27.88
N:	9	7	6	9

It can be seen from Tables 3 and 4 that for the "with feedback" condition there was a considerable decrease in the duration of both of the words measured. A Mann-Whitney U-test showed this to be significant at 0.0106 for the word "four" and significant at 0.0010 for the word "nine". In the "no feedback" condition there was no significant change.

4.3.2 **Segmental features:** although it is well known that speaker variability will be manifested in a number of prosodic characteristics of speech, there is also the vitally important aspect of what we will call "precision of articulation", and this is essentially a matter of differences arising in the individual sound segments of speech. In a study that attempts to analyse large quantities of speech automatically, this is by far the hardest task for the investigator. It is only in the last few years that computer systems capable of anything like detailed analysis of speech at the segmental level have been developed (Vaissiere 1985[19]). Such a system has been developed in the Department of Linguistics & Phonetics at Leeds University, known as LUPINS (Leeds University Phonetic INput System); the work for this was funded between 1980 and 1983 by the Joint Speech Research Unit (G.C.H.Q., award no. F7T/291/79), and since 1985 by the Alvey Programme (via S.E.R.C.: Grant no. MMI/053). The design of LUPINS is described in Roach and Roach (1983[20]).

At the time of the main body of work on the present project, the LUPINS system was in the process of being transferred on to the PDP-11/73 minicomputer system, and although it was felt to be highly desirable to analyse our speaker degradation data with the LUPINS system, it was necessary to delay this until late in 1986 when the system was judged to be running with acceptable reliability.

As mentioned above, one part of the earlier work on the present project involved playing expert human listeners extracts from the beginning and end of recordings in our corpus, and it was shown that these could be identified with reasonable accuracy. It was decided to process the same extracts with LUPINS to see if any significant differences that could be related to "precision of articulation" would be found in a statistical analysis of the results.

The final output from LUPINS consists of a string of special phonetic symbols, using a restricted alphabet that is unique to this application but is relatable to familiar phonetic categories. Each symbol has associated with it a duration figure expressed in centiseconds. The symbols represent the following categories:

VOWEL
FRICATIVE
NASAL
STOP
DIP
SILENCE

In some of our research work some of these categories are sub-divided, but for the purposes of this project it was felt preferable to use only the major category labels.

The recordings used were eight one-minute extracts taken from near the beginning and near the end of our one-hour recordings (speakers 1 - 4).

The LUPINS output files generated were processed by computer and the following questions investigated:

(i) Do mean segment type durations vary with fatigue?

(ii) Are certain segment types detected less, or more, frequently in fatigued than in un-fatigued speech? (for example, are stop consonants, which involve considerable articulatory energy, produced less frequently when the speaker is fatigued?).

The results obtained from the analysis are set out below; only the segment types Fricative, Nasal, Vowel and Stop are recorded here, since other types are difficult to interpret in short recordings. It is noticeable that fewer nasals were recognised than would be expected, which suggests a fault in the nasal detection module of LUPINS, and that some Stop average durations are considerably below what would be expected, and must therefore be considered unreliable. No consistent trend is visible in any of the results apart from a (non-significant) decrease in vowel duration in three out of four speakers (Speaker 2's result remaining constant).

TABLE 5: FREQUENCY OF OCCURRENCE AND MEAN DURATIONS OF SEGMENT TYPES

segment type	% of total segments		mean duration	
	near start	near end	near start	near end
(SPEAKER 1)				
FRICATIVE	28.4	28.4	9.6	8.4
NASAL	3.4	0.0	4.7	0.0
VOWEL	36.8	32.8	13.1	12.8
STOP	13.6	14.9	6.7	7.2
(SPEAKER 2)				
FRICATIVE	25.9	36.4	9.3	9.9
NASAL	1.9	0.0	7.0	0.0
VOWEL	35.2	30.0	14.7	14.7
STOP	9.3	18.2	5.6	6.0
(SPEAKER 3)				
FRICATIVE	43.6	41.4	20.9	18.5
NASAL	3.2	0.0	7.0	0.0
VOWEL	37.1	41.4	17.1	16.1
STOP	8.1	4.3	2.0	1.3
(SPEAKER 4)				
FRICATIVE	42.6	42.4	27.6	30.1
NASAL	0.0	0.0	0.0	0.0
VOWEL	40.4	35.6	24.2	17.3
STOP	8.1	4.3	2.0	2.5

SECTION 8: CONCLUSIONS

1. Though speakers do exhibit fatigue effects in performing speaking tasks over long periods of time, it has not proved possible in our data to identify common tendencies in these speaking changes. We believe (though we do not have the very large-scale database needed to establish this conclusively) that speaker degradation is manifested in a wide and unpredictable variety of ways. In our opinion it would not be feasible to design a speech recognition system that made modifications to word recognition templates as a function of the time for which a speaker had been speaking.

2. Some speakers were more reliable and effective than others. Again, our database was not sufficiently large to establish conclusively what factors were involved, but it appeared that two factors were important: firstly, familiarity with the equipment and technique, and secondly, professional experience in working with speech. Our conclusion is that it should be worthwhile and reasonably easy to devise a familiarisation and training scheme for future users of speech recognition equipment; this should contain a certain amount of basic instruction in Speech Science and Phonetics, as well as practical speech-input training with real-time visual feedback on performance.

3. It appears to us that the requirement for speech input from tape-recordings will within one or two years become effectively obsolete for any reasonably well-funded data-collecting activity: we understand from investigation of the latest commercially-available speech input technology that it would now be possible to equip a portable computer with a low-cost, light-weight word-recogniser facility and thus provide each operator with a fully portable, battery-powered workstation offering immediate read-out of the recognised message, and post-task downloading of stored data. This would, of course, represent a significant improvement over the system being evaluated at the outset of the present project.

4. The main achievement of the present work has been to develop and test techniques for acquiring and assessing speaker performance relevant to speaker degradation studies. We feel that it would be valuable to extend this work further by considerably enlarging the database, ensuring that it includes sufficiently clear-cut evidence of fatigue, and exploiting more fully the facility for segmenting continuous speech that we now have, in order to derive more sophisticated long-term measures of speaker performance.

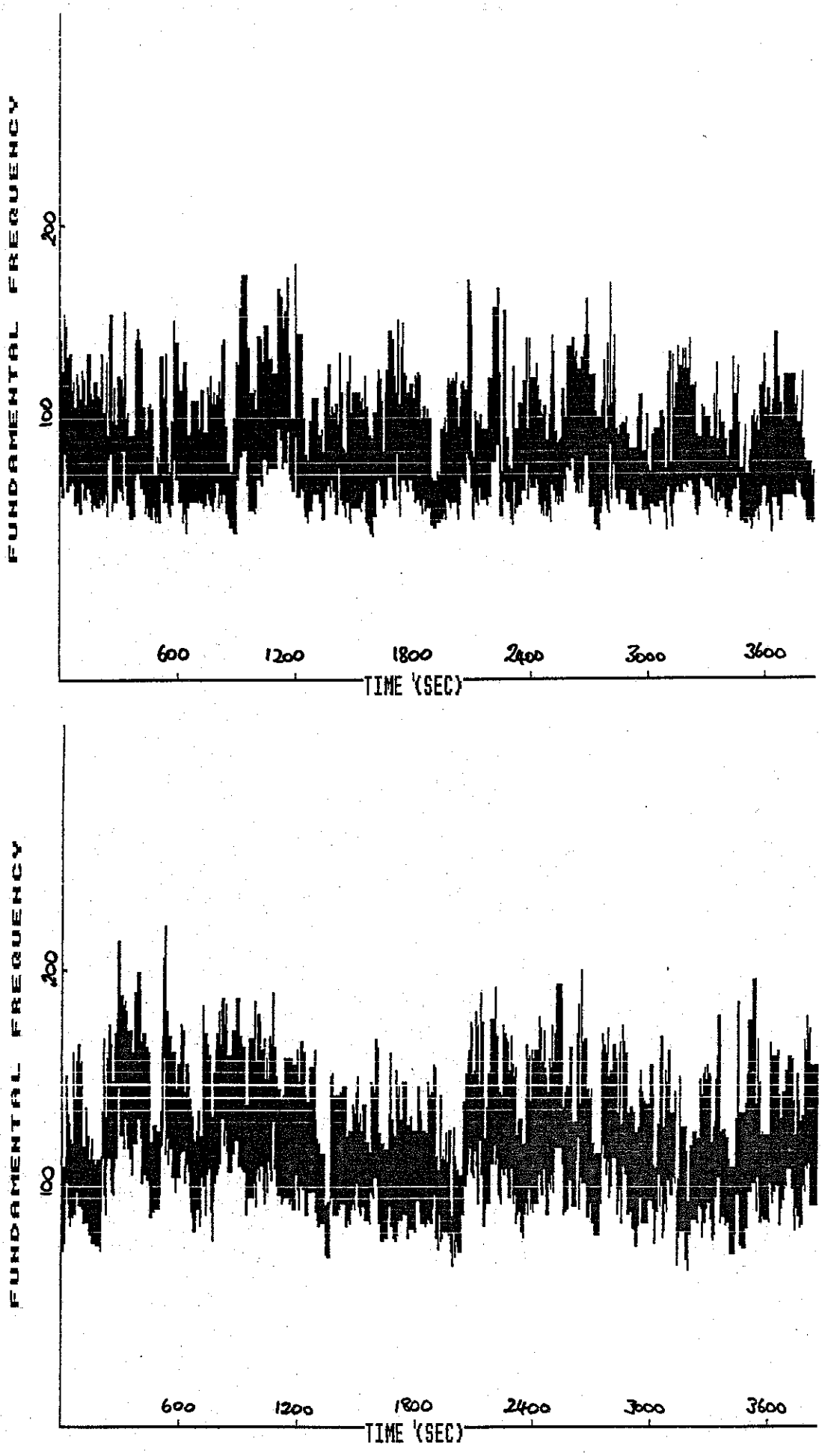

FIG.1: LONG-TERM FUNDAMENTAL FREQUENCY (Hz, +/- 1 s.d.), SPEAKER 1

Top: with feedback; bottom: without feedback.

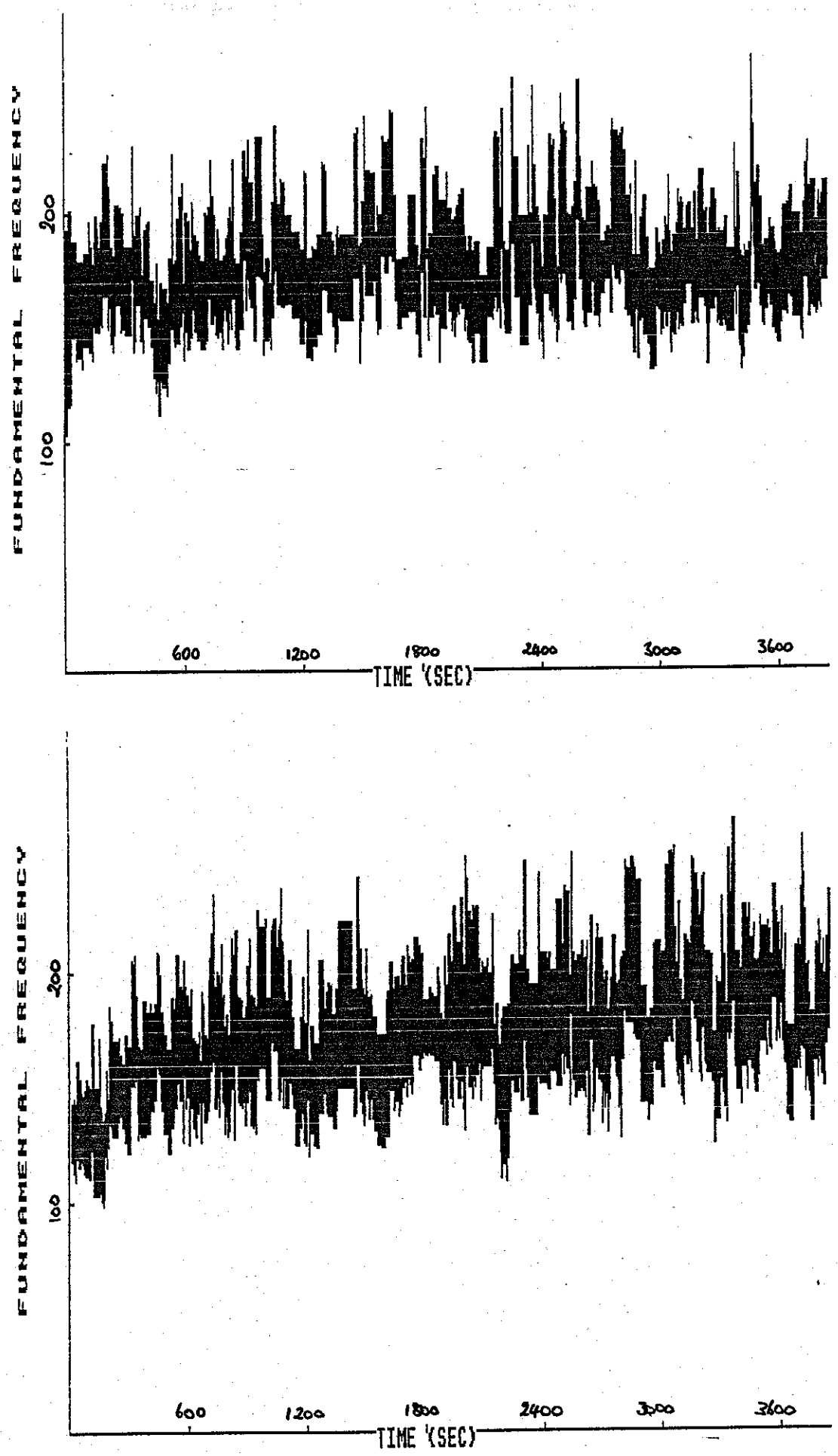

FIG.2: LONG-TERM FUNDAMENTAL FREQUENCY (Hz, +/- 1 s.d.), SPEAKER

Top: with feedback; bottom: without feedback.

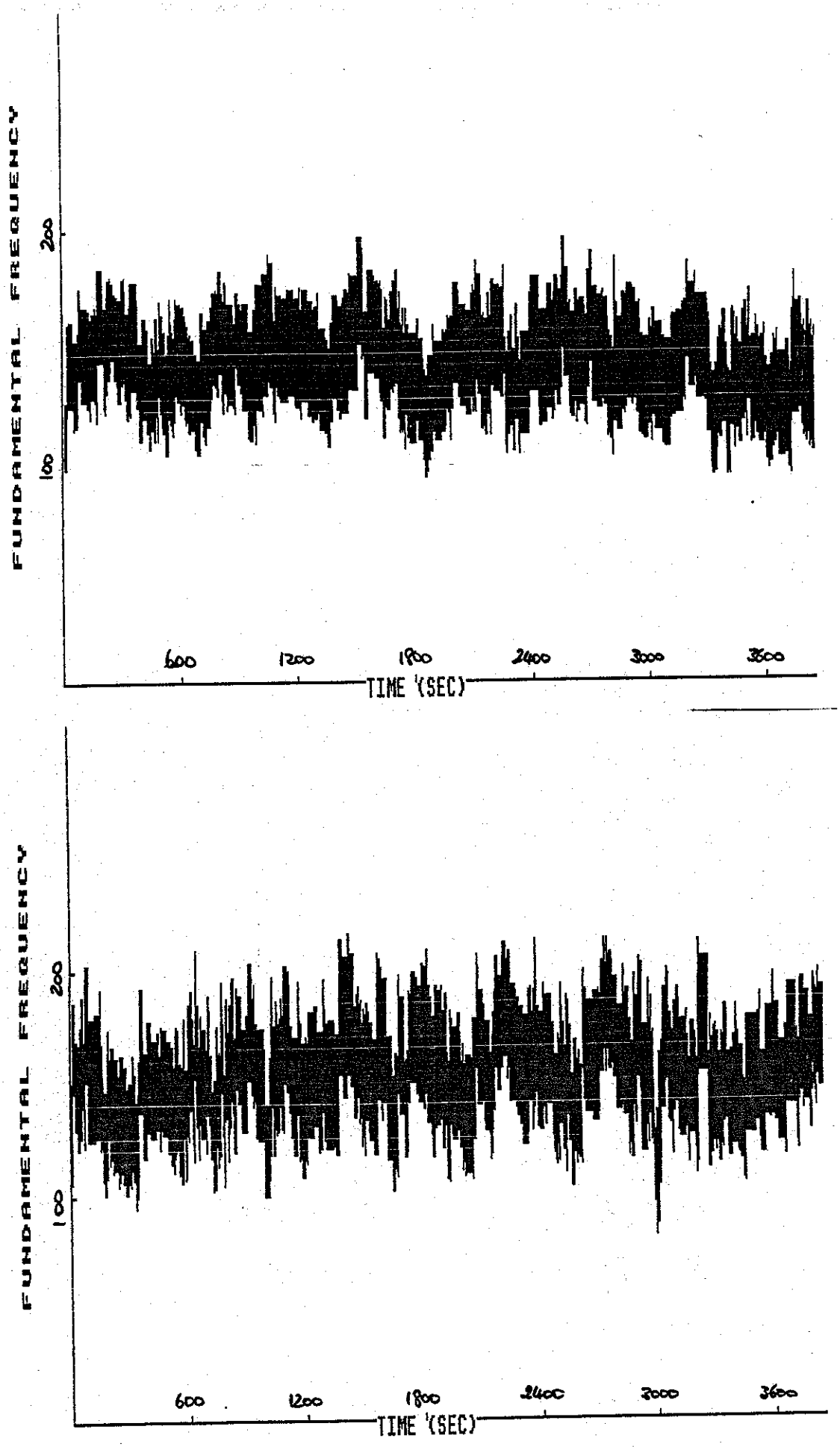

FIG.3: LONG-TERM FUNDAMENTAL FREQUENCY (Hz, +/- 1 s.d.), SPEAKER 3

Top: with feedback; bottom: without feedback.

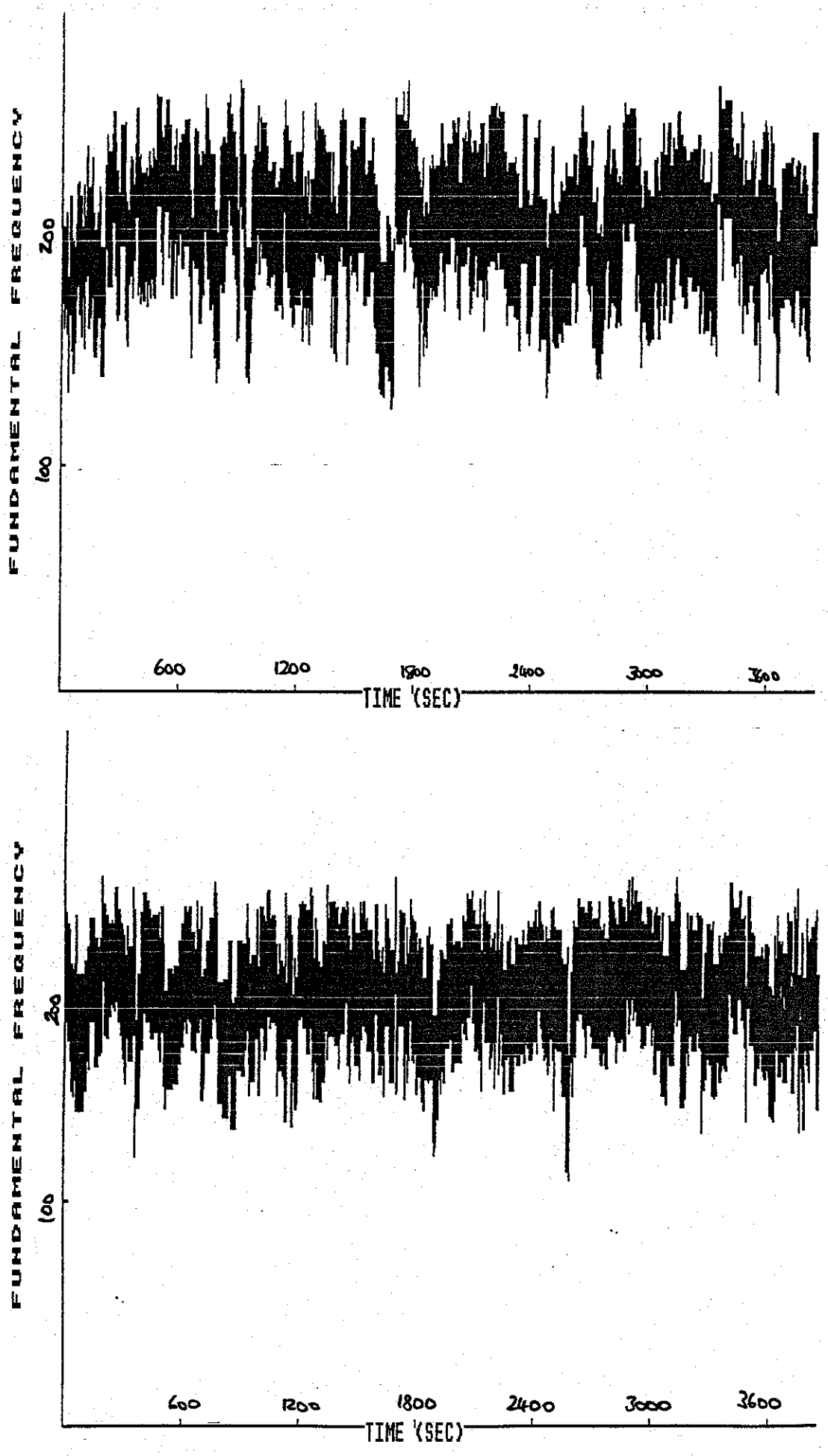

FIG.4: LONG-TERM FUNDAMENTAL FREQUENCY (Hz, +/- 1 s.d.), SPEAKER 4

Top: with feedback; bottom: without feedback.

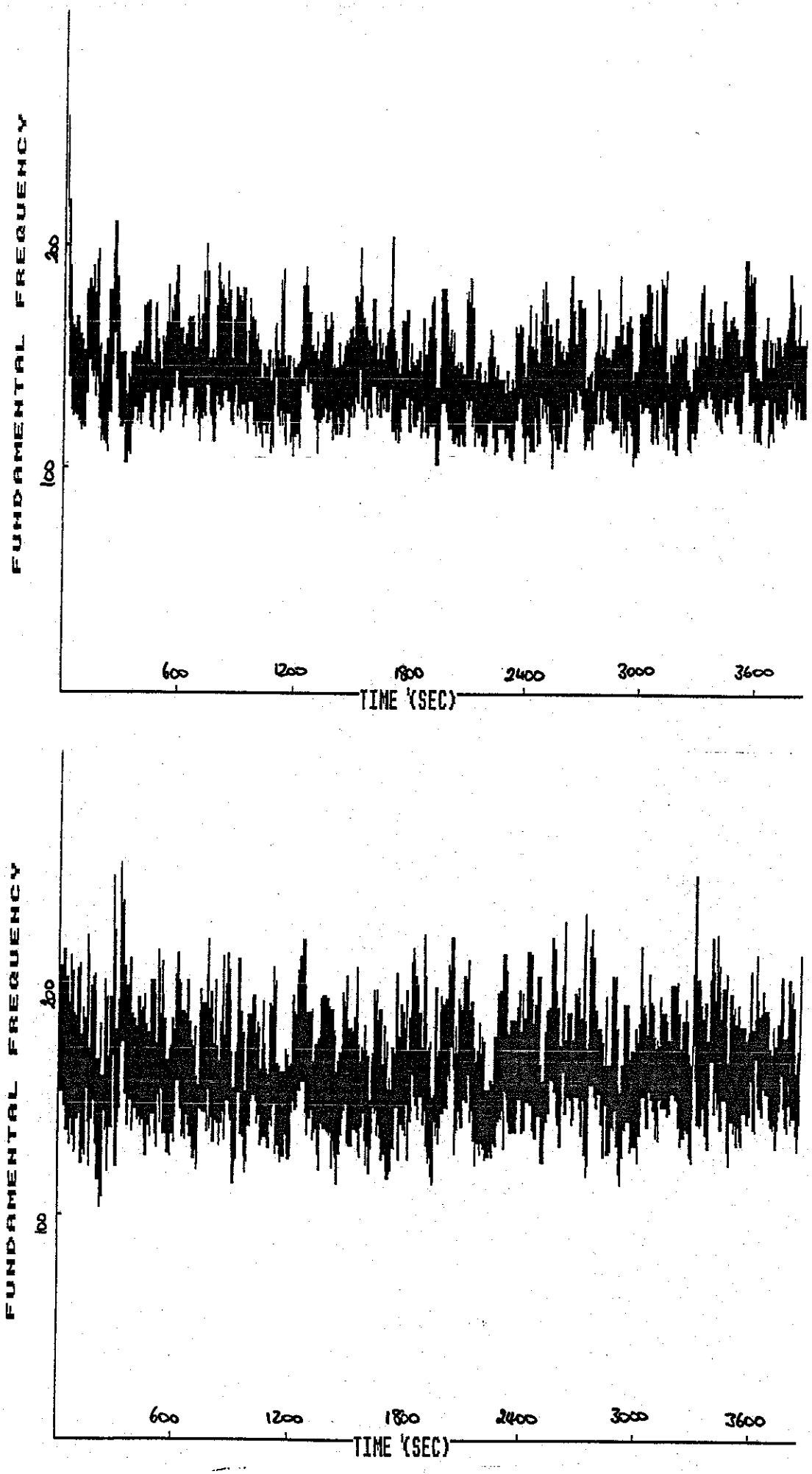

FIG.5: LONG-TERM FUNDAMENTAL FREQUENCY (Hz, +/- 1 s.d.), SPEAKER 5

Top: with feedback; bottom: without feedback.

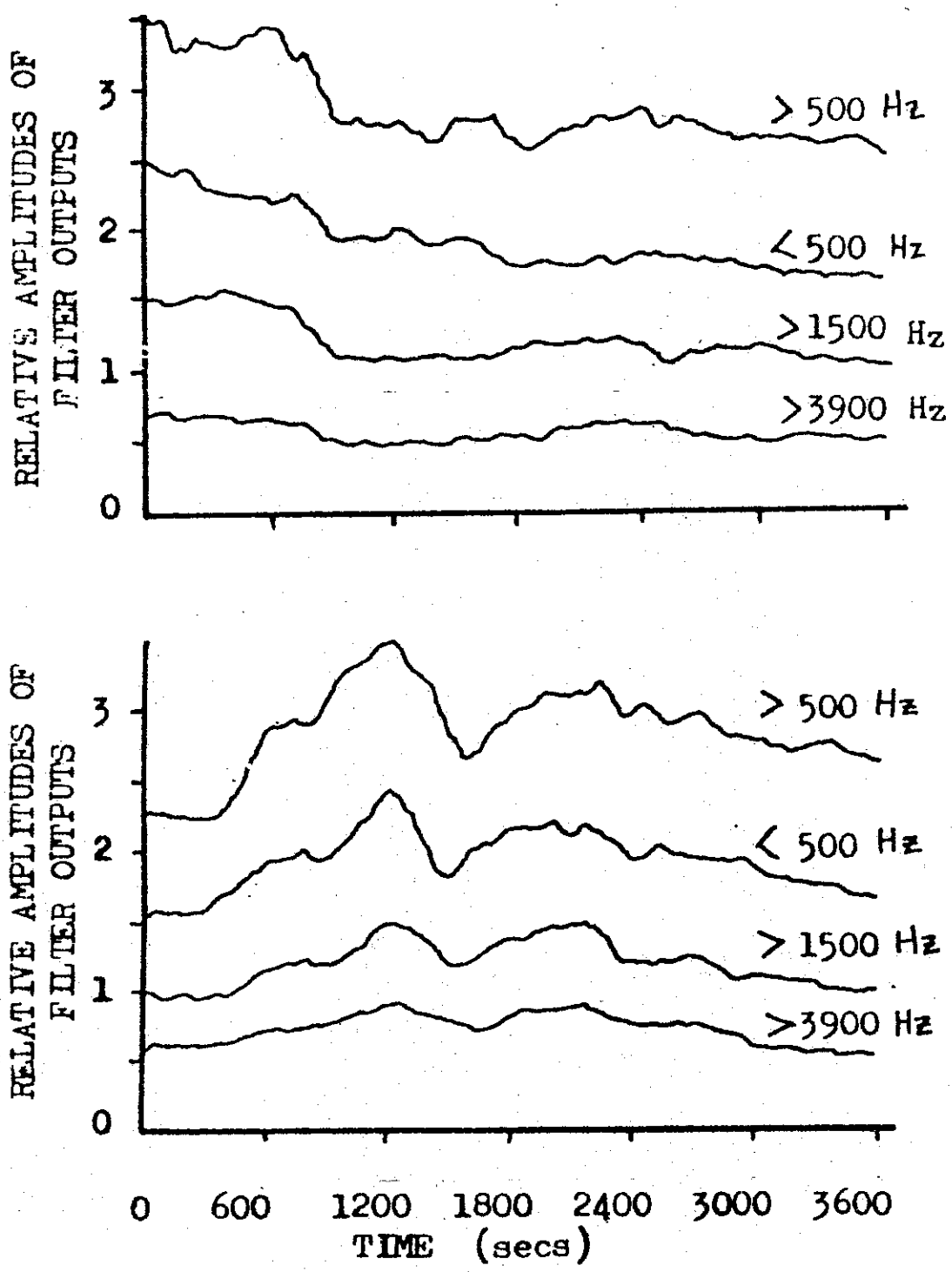

FIG.6: LONG-TERM ENERGY IN FOUR FREQUENCY REGIONS

Top: with feedback; bottom: without feedback.

8. BIBLIOGRAPHY

(i) Works cited in report.

[1] Kirby, H.R. (1983) 'The automated transcription of spoken traffic data', Technical Note 124, Inst. Transp. Stud., Univ. of Leeds (unpublished).

[2] Bonsall, P.W., Hardwick, B.A. and Kirby, H.R. (1985) 'The automatic transcription of tape-recorded data', Working Paper 205, Inst. Transp. Stud., Univ. of Leeds (paper presented at the PTRC Summer Annual Meeting, Univ. of Sussex, July 1985).

[3] Kirby, H.R. and Bonsall, P.W. (1985) 'The automated transcription of spoken traffic data', Final report to S.E.R.C., Technical Note 167, Inst. Transp. Stud., Univ. of Leeds (unpublished).

[4] Martin, T.B. and Welch, J.R. (1980) 'Practical speech recognisers and some performance effectiveness parameters', in Lea, W.A. (ed.) Trends in Speech Recognition, pp.23-38.

[5] Lea, W.A. (1980) 'Speech recognition: past, present and future', in Lea, W.A. (ed.) Trends in Speech Recognition, pp. 39-98.

[6] Wilpon, J.G. and Roberts, L.A. (1986) 'The effects of instructions and feedback on speaker consistency for automatic speech recognition', Proceedings of the IEE Conference on Speech I/O, pp. 242-247.

[7] Lim, J.S. and Oppenheim, A.V. (1979) 'Enhancement and bandwidth compression of noisy speech',Proceedings of the IEE Conference on Speech I/O, pp. 1586-1604.

[8] Roberts, L.A., Wilpon, J.G., Egan, D. and Bakk, J. (1986) 'Improving speaker consistency in an automatic speech recognition framework', Computer Speech and Language, 1.1, pp.61-93.

[9] Bobrow, D.G. and Klatt, D.H. (1968) 'BBN Report 1667', Final Report, Contract NAS, 12-138. Bolt, Beranek & Newman, Cambridge, Mass.

[10] Roach, P.J. (1983) English Phonetics and Phonology, Cambridge University Press.

[11] Dew, A.M., Hardwick, B.A., Roach, P.J., Shirt, M.A. and Kirby, H.R. (1986) 'Voice degradation problems in using automatic speech recognisers', Proceedings of the IEE Conference on Speech I/O, pp. 319-323.

[12] Steffen-Batog, M., Jassem, W. and Gruszka-Koscielak, H. (1970) 'Statistical distribution of short-term F0 values as a personal voice characteristic', in Jassem, W. (ed.) Speech Analysis and Synthesis 2, 195-206. Warsaw, Polish Academy of Science.

[13] Hiller, S., Laver, J. and Mackenzie, J. (1984) 'Durational aspects of long-term measurements of fundamental frequency perturbations in connected speech', Work in Progress,17, pp. 59-76, Edinburgh University Dept. of Linguistics.

[14] Fourcin, A.J. and Abberton, E. (1972) 'First applications of a new laryngograph', J. Medical & Biol. Illust., 21, pp. 172-182.

[15] Leon, P.R. and Martin, P. (1972) 'Machines and measurements', in Bolinger, D. (ed.) Intonation, pp.30-47; Penguin.

[16] Johns-Lewis, C. (1986) 'digital analysis of pitch and silence in three speech styles', Proceedings of the IEE Conference on Speech I/O, pp. 281-286.

[17] Vilkman, E. and Manninen, O. (1986) 'Changes in prosodic features of speech due to environmental factors', Speech Communication, 5, pp. 331-345.

[18] Lehiste, E. (1970) Suprasegmentals, M.I.T.

[19] Vaissiere, J. (1985) 'Speech recognition: a tutorial', in Fallside, F. and Woods, W.A. (eds.) Computer Speech Processing, pp. 191-242.

[20] Roach, H.N. and Roach, P.J. (1983) 'Automatic Segmentation and labelling of speech sounds from different languages', Working Papers in Linguistics & Phonetics, 1, pp.91-95.

(ii) Papers produced during the project.

Dew, A.M., Hardwick, B.A., Roach, P.J., Shirt, M.A. and Kirby, H.R. (1986) 'Voice degradation problems in using automatic speech recognisers', Proceedings of the IEE Conference on Speech I/O, pp. 319-323.

Dew, A.M., Hardwick, B.A. and Roach, P.J. (1985) 'Measurement of changes in voice and speech during long recording sessions', Technical Note 177, Inst. Transp. Stud., Unive. Leeds.

Hardwick, B.A., Bonsall, P.W. and Kirby, H.R. (1986) 'Applied evaluation of speech recognisers with respect to tape-recorded data', 65th. Annual Meeting Transptn. Resch. Bd., Washington DC, January 1986 (also available as: Work.Ppr. 213, Inst. Transp. Stud., Univ. of Leeds.

Kirby, H.R. (1986) 'How can speech recognisers help applied research in the Civil Engineering, Transport and related industries?'; notes of a seminar held at the University of Leeds on 5th. November 1986. Tech. Note 203, Inst. Transp. Stud., Univ. of Leeds.

Kirby, H.R. (1986) Report on visit to USA on speech recognition in Jan. 1986. Report to the U.S. Army under Contract no. DAJA45-86-M-0144.

Kirby, H.R. and Bonsall, P.W. (1985) 'THe automatic transcription of spoken traffic data'. Final report to S.E.R.C.; Tech. Note 167, Inst. Transp. Stud., Univ. of Leeds.

Kirby, H.R. and Roach, P.J. (1985) 'Voice degradation in using speech recognisers for transcribing inventory data'. Project Description; Tech. Note 166, Inst. Transp. Stud., Univ. of Leeds.

Kirby, H.R. and Roach, P.J. (1985) 'Voice degradation in using speech recognisers for transcribing inventory data'. First periodic report; Tech. Note 178, Inst. Transp. Stud., Univ. of Leeds.