TIME-STRUCTURE MEASURES OF VOCAL FOLD VIBRATION

Maidment J.A., Howard D.M., and Smith, D.A.J.

Dept of Phonetics and Linguistics, University College London, UK.

1. INTRODUCTION

This paper describes two measures of the gross time-structure of laryngeal activity during speech. The measures were first mentioned in [1] in connection with the comparison and evaluation of pitch estimation devices. However, at that time the work was still at a preliminary stage. The work reported here is concerned with the development and verification of the technique and has three major aims:

(i) the verification of input and analysis procedures
(ii) the estimation of optimum input sample size
(iii) the description of salient features of the measures

The first of the measures, Sx, is the probability-density function of the durations of periods when vocal fold vibration is absent. In Fig. 1 (a), which shows annotated speech pressure (SP) and laryngograph (Lx) waveforms for the utterance [apapa], together with estimates of the durations of laryngeal period (Tx) derived from the Lx waveform [2] , the two periods of laryngeal silence are marked as Sx1 and Sx2. These correspond to the two intervocalic plosive segments. In general, the interruption of phonation during speech is caused by one of two events. First, a pause on the part of the speaker will be marked by the cessation of vocal fold vibration. Secondly, the production of many obstruent consonant segments, whether phonologically voiced or voiceless in the language concerned, is also likely to cause a break in phonation. In Fig. 1 (b), for example, the period of laryngeal silence, marked as Sx1, corresponds to the voiced plosive segment [b:] in the utterance [wamab:a].

The second measure, Vx, is the probability-density function of the durations of uninterrupted periods of vocal fold vibration. In Fig. 1 (a) there are three such periods, marked Vx1 - 3, corresponding to the vowel segments, while in Fig. 1 (b) there are two, the first, Vx1, corresponding to the sonorant sequence [wama] and the second, Vx2, to the final vowel.

2. INPUT AND ANALYSIS

Both Sx and Vx are computed from a Tx representation of speech (see Figs. 1 (a) and 1(b)), using a suite of programs written for a BBC microcomputer. Typically, Tx is available in real time from the output of a layryngograph processor [2] or are estimated from the speech pressure waveform by one of several means (see [3] for a review). Tx values are input to the micro where they are stored to 15 bit accuracy. The Tx values below 32 ms are stored in microseconds and values above 32 ms are stored in milliseconds. The second case is signalled by setting the most significant bit of the two-byte word. When the input is terminated by the user or the microcomputer's memory is filled, the file of Tx values is written to floppy disk for further analysis.

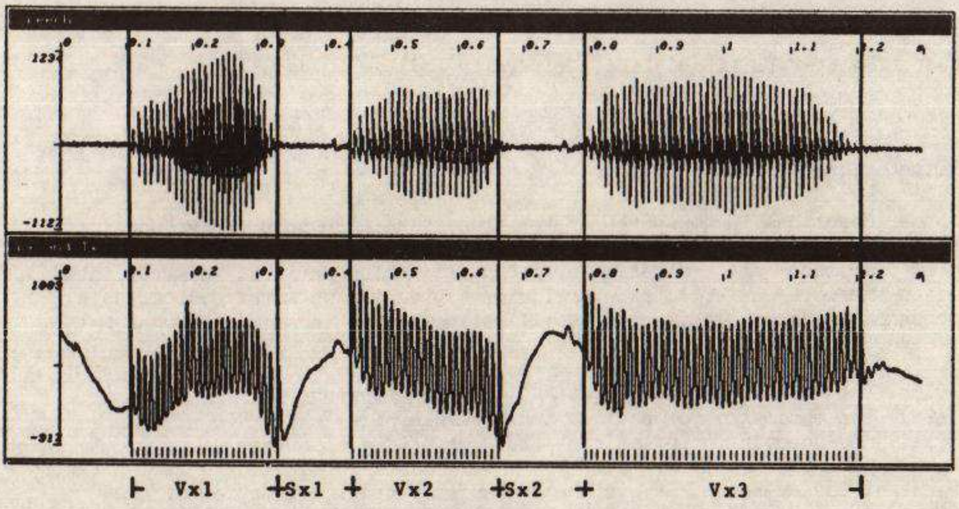


Figure 1a: Speech pressure, laryngograph and Tx waveforms for [ɑpɑpɑ] spoken by a normal adult male

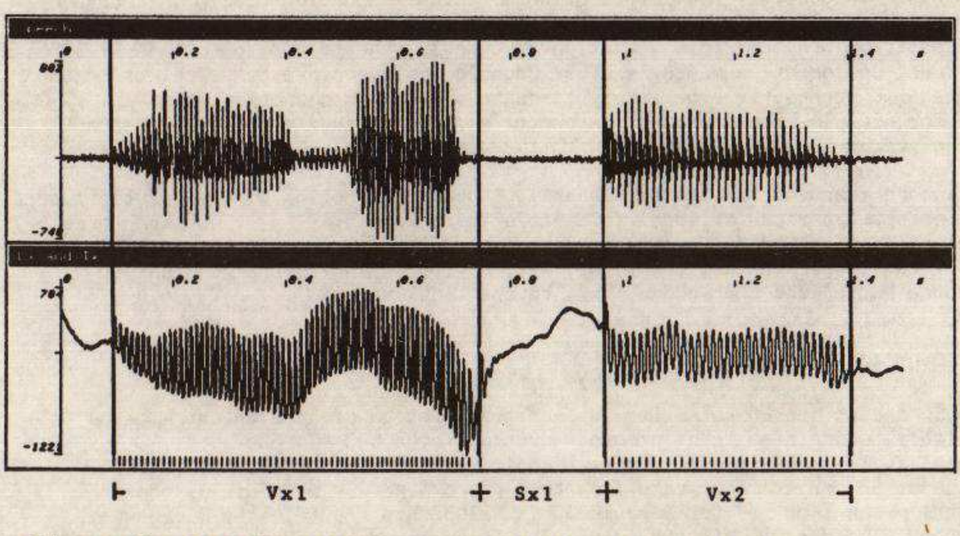


Figure 1b: Speech pressure, laryngograph and Tx waveforms for [wɑmɑb:ɑ] spoken by a normal adult male

A maximum of 12544 Tx samples can be stored and this corresponds, on average, to about 2.5 - 3 minutes of speech for a male speaker and 2 - 2.5 minutes for a female.

The Vx and Sx analysis programs can deal with multiple Tx file input and therefore are capable of producing an analysis of an indefinitely long speech sample. Both analyses result in a histogram of percentage frequency of occurrence as a function of quantised duration. In the case of Sx, the duration concerned is simply that of Tx values between 32 ms (the threshold value for voicing) and 32 secs. This range is divided into 128 logarithmically related intervals. Vx displays the summed durations of Tx occurring between two successive above-threshold values. The range of durations for this histogram is 1 ms - 10 seconds and this again is divided into 128 logarithmically related intervals. The vertical scale of both displays is also logarithmic. The analysis procedures compute a variety of statistics of the distribution, including mean and standard deviation. Examples of Sx and Vx distributions for two male speakers may be found in Figs 2(a) and 2(b). These were the result of analysing recordings of a reading lasting approximately 16 minutes.

3. VERIFICATION OF INPUT PROCEDURE

As has been explained, the Tx values are stored in the computer in two forms indicated by bit 16 of the data word. To test the input procedure of the system it was necessary to apply two pulse trains to the computer, one with a fundamental period above and one below the 32 ms breakpoint. The frequencies applied were 1100 Hz and 15 Hz which gave values in the Tx data files of 038Ch and 8043h, corresponding to 908 microsec and 67 ms respectively (1101 Hz and 14.92 Hz).

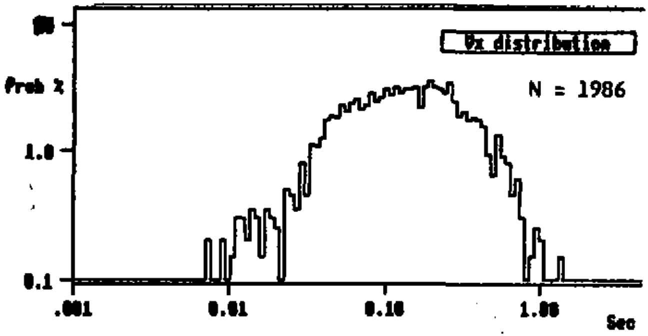
4. VERIFICATION OF ANALYSES

To test the analyses, a dummy Tx data file was constructed containing known Sx and Vx intervals. The data file was filled using a linear progression of y=x which produces a distribution showing clearly the effect of the logarithmic probability scale as seen in Figs. 3(a) and 3(b)

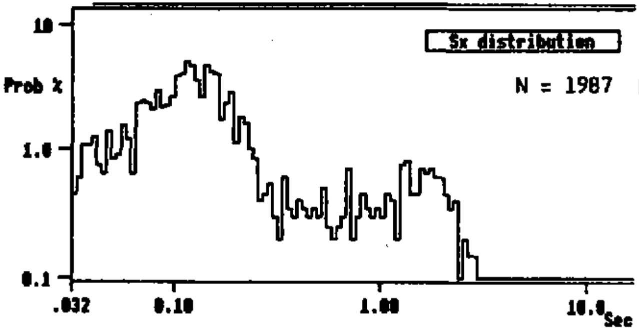
5. ESTIMATION OF OPTIMAL INPUT SAMPLE SIZE

Fig. 1 shows that the phonetic structure of the input text will have a major effect on the final form of both Sx and Vx analyses. The question arises whether it is possible find, for a particular style of speech (free informal conversation, lecturing, reading and the like), a size of input sample at which the distributions of Sx and Vx become stable in the sense that further input will cause only minimal alterations in their shape. This could then be recommended as the minimum input sample.

In order to discover if such a minimum input sample exists for Sx and Vx and to attempt to estimate its size a procedure was followed which was analogous to

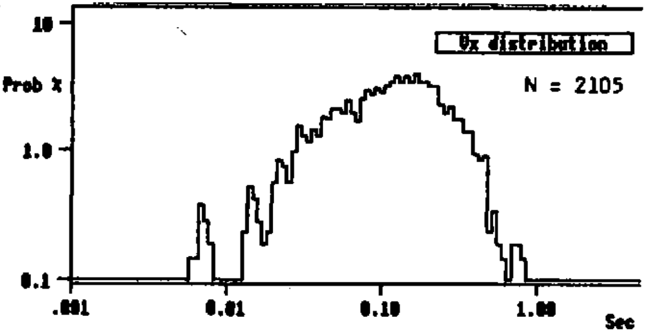


Vx distribution for speaker JM
16 minute sample duration.

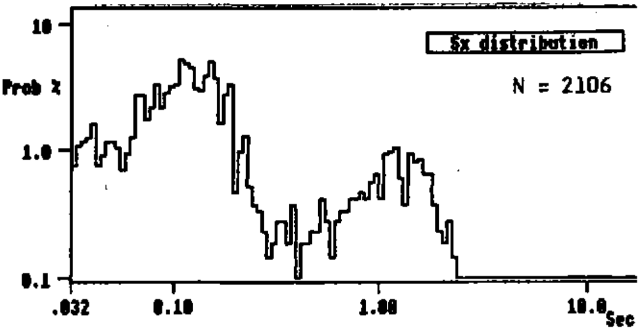


Sx distribution for speaker JM
16 minute sample duration.

Figure 2a

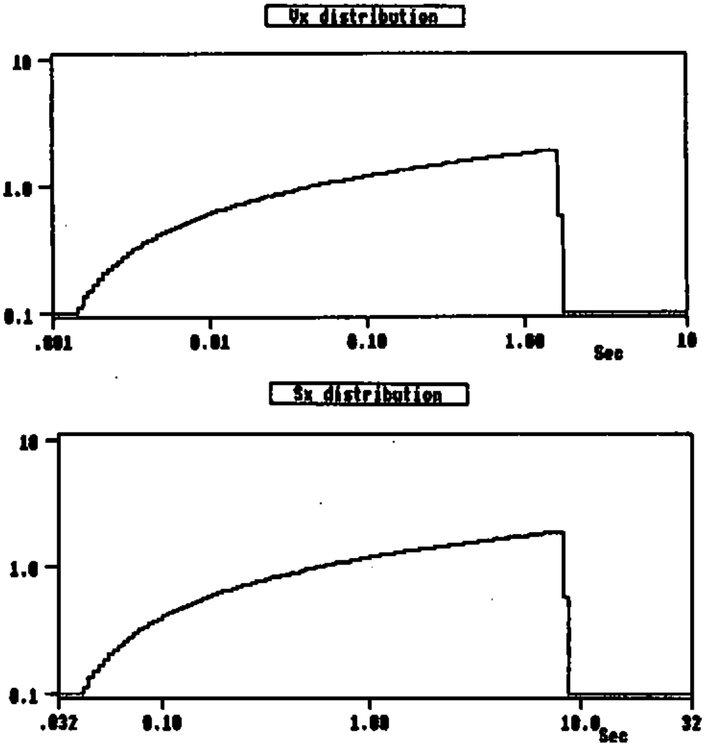


Vx distribution for speaker DH
16 minute sample duration



Sx distribution for speaker DH
16 minute sample duration.

Figure 2b

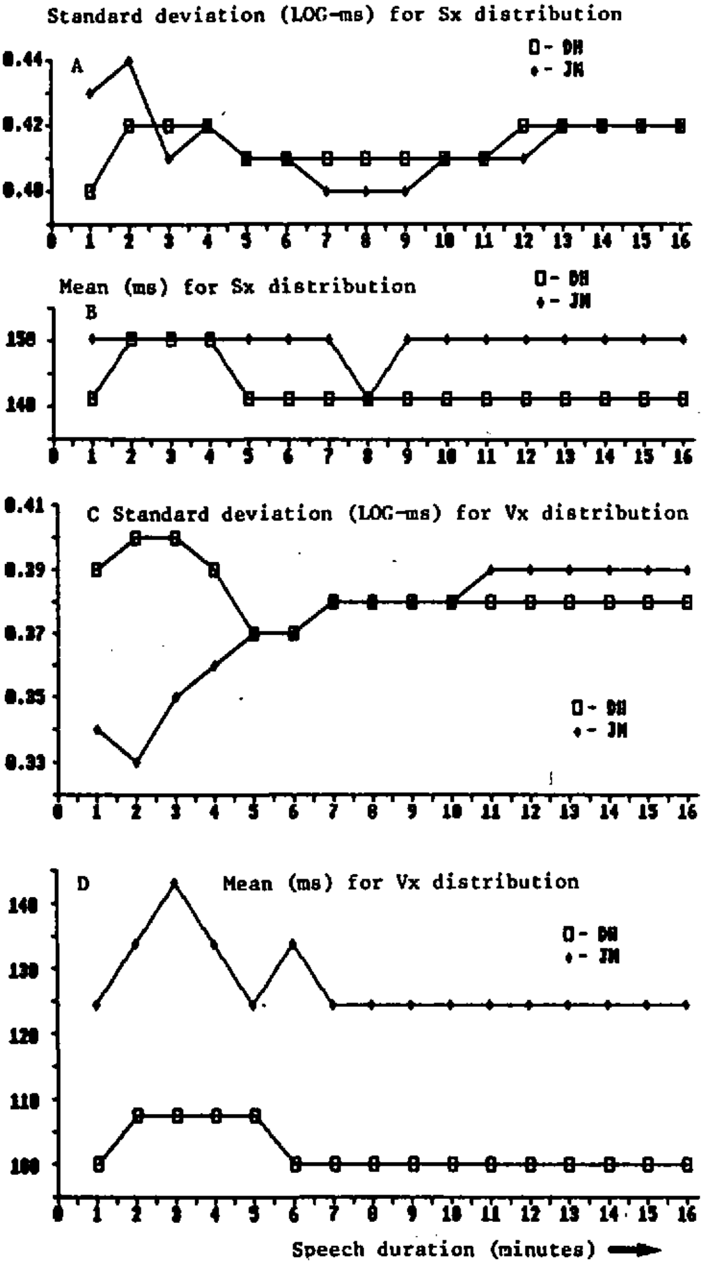


Figures 3a & 3b: Vx and Sx plots for linear interval data.

that used in [4] to estimate the stability of speech fundamental period distributions. The recordings used to produce the distributions in Fig. 2 were divided into approximately 1 minute segments and Sx and Vx distributions were produced for the first of these segments, then the first and second combined and so on. At each stage the mean and standard deviation of the distribution was calculated. The results are are presented in Fig. 4.

Fig 4(b) shows that the mean Sx value for each speaker settles down to a constant value - after 11 minutes for speaker JM and after 7 minutes for speaker DH. The corresponding sample sizes can be found in the tables relating sample size to duration. They are 1385 and 1010 samples respectively. The standard deviations for Sx for the two speakers are plotted against increasing sample duration in Fig. 4(a). Both functions settle to a constant figure after 13 minutes, although there is only minimal change after 10 minutes. The sample size at 13 minutes is 1611 samples for JM and 1759 for DH.

Figure 4: Graphs of mean and standard deviation changes in Sx and Vx.



Standard deviation (LOG-ms) for Sx distribution

Mean (ms) for Sx distribution

C Standard deviation (LOG-ms) for Vx distribution

Mean (ms) for Vx distribution

Sx sample size		
Sample duration (min)	JM	DH
1	128	139
2	240	264
3	364	402
4	482	546
5	633	713
6	757	863
7	982	1010
8	1035	1147
9	1156	1278
10	1265	1420
11	1385	1487
12	1512	1621
13	1611	1759
14	1745	1900
15	1868	2040
16	1987	2106

Vx sample size		
Sample duration (min)	JM	DH
1	128	139
2	240	264
3	363	401
4	480	545
5	631	712
6	755	862
7	980	1010
8	1033	1147
9	1155	1277
10	1264	1419
11	1384	1486
12	1511	1620
13	1610	1758
14	1744	1899
15	1859	2039
16	1986	2105

Turning now to Vx, Figs. 4(c) and 4(d) show that the mean and standard deviation for both speakers are constant after 11 minutes of speech input. The sample sizes are 1384 samples for JM and 1486 for DH.

It would appear then that an input duration of 10 - 11 minutes of speech which produces a sample size of around 1400 might tentatively be taken as a minimum input for these distributions.

It must be emphasised that the above findings are only necessarily valid for the style of speech investigated, which may be characterised as unprepared, reasonably fluent reading at moderate tempo.

6. GENERAL CHARACTERISTICS OF Sx AND Vx

Figs. 2(a) and 2(b) show that Sx for both speaker DH and speaker JM is clearly bimodal for the style of speech investigated. The lower values mode is dominant in both cases. One explanation for this feature is that it reflects the two differing causes of phonation breaks mentioned above. The concentration of lower values may be due to the occurrence of obstruent segments, while the concentration of higher values is probably caused by pauses or hesitations by the speaker. Such pauses are relatively infrequent in this style of speech.

The Vx is essentially unimodal for both speakers and this reflects the fact that a Vx period has only one phonetic correlate, that is an uninterrupted sequence of sonorant segments.

7. CONCLUSIONS

While it would be inaccurate from both the linguistic and acoustic viewpoints to regard the larynx as a simple two-state device, it is clear that the alternating presence and absence of vibratory activity in the larynx and the consequent periodic excitation of the vocal tract or lack thereof is a basic feature of normal human speech. It is therefore to be expected that many features of the speech situation may be reflected in the time-structure of vocal fold vibration.

This investigation has established, at least tentatively, that for one specific style of speech, the distributions of the two measures of time-structure become stable with an input sample of manageable size. There is a need, however, to investigate how the measures behave when aspects of the speech situation are varied. These include speech tempo, formality of speech, sex of speaker and speaker accent. Also it is likely that many types of structural or functional abnormality of the larynx will have consequences for the shape of both Sx and Vx distributions. The usefulness of these measures as part of a diagnostic procedure together with other analyses of larynx activity (see [5]) needs to be investigated.

ACKNOWLEDGMENT: We would like to thank the author of the Tx Processing System for permission to publish the distribution figures in this paper.

8. REFERENCES

1. Howard D.M., Maidment J.A., Smith D.A.J. & Howard I. (1986) Towards a comprehensive quantitative assessment of the operation of real-time fundamental frequency extractors. IEE Conference Publication No. 258, 172 - 177.

2. Fourcin A.J. & Abberton E.R.M. (1971). First applications of a new laryngograph. Med. and Biol. Illust. 21, 172 - 182.

3. Hess W. (1983) Pitch determination of speech signals., Springer-Verlag, Berlin.

4. Mead K. (1974) Identification of speakers from fundamental frequency contours in conversational speech. JSRU Report No. 1002

5. Fourcin A.J. (1981) Laryngographic assessment of phonatory function. ASHA Reports 11, 116 - 127.