Article:

Howard, David Martin orcid.org/0000-0001-9516-9551, Maidment, John, Smith, David et al. (1 more author) (1986) Towards a comprehensive quantitative assessment of the operation of real-time fundamental frequency extractors. IEE Conference Publication 258. pp. 172-177.

TOWARDS A COMPREHENSIVE QUANTITATIVE ASSESSMENT OF THE OPERATION OF REAL-TIME FUNDAMENTAL FREQUENCY EXTRACTORS

DAVID M HOWARD, JOHN A MAIDMENT, DAVID A J SMITH and IAN S HOWARD

PHONETICS and LINGUISTICS DEPARTMENT, UNIVERSITY COLLEGE LONDON, UK

ABSTRACT

Reliable measurement of speech fundamental frequency is an essential element in many aspects of research. There are many methods available for such measurement, but there is no rigorous technique which allows a quantitative evaluation of these methods.

This paper discusses work being carried out at UCL to investigate the possibility of providing a viable methodology for device assessment. Three acoustically based devices are used alongside a standard (the laryngograph) to illustrate progress to date.

INTRODUCTION

Phoneticians, linguists, speech therapists and musicologists are members of just some of the professions whose work can depend upon a reliable estimation of speech fundamental frequency (Fx). The design of devices and algorithms to extract Fx from a speech input has been, and still is, a prime area in speech research. Such a device should isolate those portions of the input speech which are neither voiceless nor silent, and determine their periodicity. Many algorithms, (1) gives an extensive review, utilising properties of a periodic signal in the time domain and/or the frequency domain, have been proposed which attempt to do this, but to date, no Fx estimation algorithm or device exists which operates reliably for all speakers in all possible speech environmental conditions. Those algorithms which are used, have typically been designed for a particular application, and in most cases, the chosen technique has to undergo an elaborate optimisation procedure.

This optimisation process during device development is most time consuming, principally because there is no quantitative method with which the operation of a device can be quickly evaluated, to ascertain for example, whether altering a particular operating parameter improves or worsens overall device performance. One may, for example, alter one parameter to improve the detection of the start of voiced segments, but that change might in itself mean that the ends of voiced segments are no longer so reliably defined. It is also the case that many professional users have to rely heavily on the sometimes rather insubstantial claims made regarding the ability of a particular device to estimate Fx. Such a user would be disadvantaged because there is no quantitative benchmark against which to estimate, say, how suitable a given system, which will have been designed for a particular situation, might be for another user´s intended application.

There is a paucity of work comparing the operation of such devices. The most extensive is (2), which involved a heavy interactive human workload and would thus not be suitable for making quick checks during device development. This paper discusses various measures, some new and some already routine, which are being carried out on the outputs from, at present, four Fx devices, one of which can justifiably be considered as a ´standard´ against which the operation of others can be assessed. These measures have found and indeed will find successful application in the quantification of normal, pathological and synthetic voice production, after the appropriate choice of a speech fundamental frequency estimation device has been made.

DEVICES USED IN THIS STUDY

Devices designed to estimate Fx can be divided into the following four classes: frequency domain devices, which rely on the fact that in voiced speech the spectrum is essentially harmonic; time domain devices, which rely on the fact that voiced speech is essentially periodic; hybrid devices, which combine various features of time and frequency domain devices; and devices which derive their input directly from the larynx.

At UCL, many years of experience have been gained with the laryngograph (3) which fits into the last-named category. This device derives a fundamental period measure directly from the source of voiced sounds - the vocal folds. Hence the laryngograph output waveform (Lx) provides the basis for a rigorous indication of Fx and it is used as the ´standard´ against which other devices can be compared, a practice endorsed in (4).

In the tests, three acoustically based devices are currently tested against the laryngograph. Two are widely known and well established methods: the cepstrum method (5), which operates in the frequency domain; and the Gold/Rabiner algorithm (6), which operates in the time domain. The third is a time domain peak-picking device (7), which is a small pocket-sized battery-powered system that has been developed as part of the EPI group cochlear implant prosthesis (8). Each of these acoustically based devices exist in two forms: as a real-time hardware device, used in the tests involving passages of text; and as software implementations on a Masscomp 5500 system which have been developed as part of the Alvey programme of the speech pattern algorithmic representations (SPAR) group, used for the detailed study of device output waveforms for short speech input segments.

DEVELOPMENT OF A DEVICE ASSESSMENT STRATEGY

The errors which are made by Fx estimation devices have been itemised in (2) in four categories: a) gross; and b) fine pitch determination errors; c) voiced-to-voiceless errors; and d) voiceless-to-voiced errors. These errors are all concerned with the fine detail of algorithm operation, and would appear to encompass the essential elements required to carry out device assessment. Whilst the terms 'voiced' and 'voiceless' are used to describe a phonetic opposition, and it is of prime importance to investigate how the devices cope with the transition from one to the other, it should be noted that there are occasions when the vocal folds do not vibrate but the sound would be 'voiced' and perceived to have a pitch, as for example, in whispered speech. None of the present day Fx estimation devices will cope with such a situation without the addition of some speech recognition resources.

In order to obtain some useful quantitative measure of the operation of these devices in these terms, it is essential to be able not only to measure these parameters in a useful manner, but also to present the results in a fashion which makes for ease in their interpretation. In an attempt to achieve this in this study the outputs from the four devices are being investigated and compared at two levels: a "macro" level, in the sense that a comparison is made using a complete 2-3 minute passage of spoken text as input; and a "micro" level where a detailed inspection of the device output waveforms obtained is made when a short input is used. At the macro level, statistical procedures already exist (see below) which have been developed for the quantification of normal and pathological, for example see (9), and synthetic (10) voice production parameters, and these measures are used as the basis for the development of new procedures specifically for this study.

It is hoped that in the future these new micro and macro measures will begin to be combined in such a manner that the micro method time-aligns the device output waveforms for "best-fit", thus providing a fixed time axis to allow the macro measure properly to begin to quantify the categories a) to d) listed above. Thus in this initially rather simple approach, the operation of devices can be ordered by merit against the standard, or progress during device development can be quantitatively monitored.

MEASURES AT THE 'MACRO' (WHOLE PASSAGE INPUT) LEVEL

The laryngograph output waveform (Lx) is used as a benchmark in the assessment. This waveform is derived by measurement of the varying electrical impedance between two electrodes placed externally on either side of the speaker's larynx (3). The appropriately polarised Lx waveform gives a direct measure of vocal fold contact area with time, thus defining the point when the vocal tract is acoustically excited with each vocal fold closure very clearly (see figure 1b). In order to make use of Lx for device comparison, a clear indication of each instant of closure is required from which a "standard" fundamental frequency (Fx) or fundamental period (Tx) measure can be defined. Typically a Voiscope (9) is used, but in the cases plotted below a Masscomp 5500 implementation has been used, which generates a pulse at each point of closure to allow Tx (see figure 1c) and the corresponding Fx (see figure 2a) to be measured on a period by period basis.

Given a digital representation of the larynx period values for a sizeable passage of speech, a number of summarizing analysis techniques may be applied. Perhaps the most obvious of these is the computation of the probability-density function of the Tx values. The practice at UCL has been to compute the percentage normalized frequency of occurrence of Fx values derived from Tx and quantized to 128 logarithmically equal intervals in the range 30.52 to 1000Hz. The results of such analysis, called Dx, are displayed as a histogram with log scales for both horizontal and vertical axes (see Figure 3). The software package which performs this analysis is implemented on a BBC micro computer system and contains options for the analysis to proceed on single Tx values (1st Order), doublets of successive Tx values (2nd Order) or triplets of successive Tx values (3rd Order). There is also an option to compute various summary statistics of the distribution, such as mean, mode, median, variance and estimates of the range. For the purposes of comparing one Dx distribution with another, both visually and statistically, it has been found useful to plot Dx cumulatively (see Figure 4). The measure of similarity of distributions is the Kolmogorov-Smirnov (KS) statistic, which is a measure of goodness-of-fit. This was chosen for two main reasons: ease of computation (the KS statistic is simply the maximum absolute difference between two cumulative step functions) and the fact that no assumption of an underlying Gaussian population distribution must be made.

The second type of 'macro' analysis which may be applied is the computation of the probability-density of first order Tx transitions. The Tx values are first converted to Fx and quantized to 64 logarithmically equal intervals in the range 30.52 to 1000Hz.The distribution which results from this analysis is called Cx and is illustrated in Figure 5. The probability of transition between any pair of quantized Fx values is indicated by the darkness of the marking at the relevant co-ordinates of the diagram.

The above types of analysis are of fairly long standing. Two other analyses of Tx have been recently developed specifically for the purposes of device comparison, although it is envisaged that these too will find wider applications in speech research. In view of the classification of device errors given in (2) mentioned above which includes voiced-to-voiceless errors and voiceless-to-voiced errors, it was thought that it might be fruitful to investigate the distributions of durations of laryngeal silence and the durations of uninterrupted laryngeal activity and to compare the output of the various devices under these two types of analysis. Thus, the two new analyses called Sx and Vx show the probability-density function of silent periods and voiced periods in the output of the devices. The threshold value for a break in voicing is a period duration exceeding 32.77 ms. Sx, therefore is simply the distribution of Tx values between 32.77 ms and 32.77 s, which is the maximum value which can be stored by the input routine. Vx is the distribution of of the sums of Tx periods occurring between successive non-voiced portions of speech. Sx and Vx distributions may be found in Figures 6.

MEASURES AT THE 'MICRO' (SINGLE PHONE INPUT) LEVEL

The measures which are currently being investigated at the micro level involve detailed measurements on the output Tx waveforms (one pulse per voiced speech period) from the devices. In these initial stages, short input speech pressure and Lx waveforms are employed, and this discussion will be restricted to an isolated citation form vowel [ɑ], as in 'far', spoken by a normal male.

Each of the three acoustically based algorithms, implemented digitally on a Masscomp 5500 system, takes a speech pressure waveform sampled at 12.8kHz as input, and produces a Tx waveform as output. The Lx waveform is also sampled at 12.8kHz, and this is digitally processed to produce a Tx waveform which is used as the 'standard'. In all the Tx waveforms, a pulse consists of a single non-zero value with all other values being zero. The Tx waveforms obtained from all four devices are shown in figure 1, along with the original speech pressure and Lx waveforms. These Tx waveforms can be transformed to Fx contours on a period by period basis without smoothing, see figure 2, to give a clearer visual impression of the device outputs, indeed, this method has been used to make an initial device comparison in the past (11). In this case it can be seen that the rise-fall intonation pattern is clear in each output, although a closer inspection clearly reveals differences at the 'micro' level.

The current work at the micro level involves using the four Tx waveforms as the input to a program which correlates the standard Tx waveform with each of the test Tx waveforms in turn. A correlation array is obtained on a point by point basis by delaying the test Tx waveform with respect to the standard waveform and then multiplying them together. Then the test Tx waveform is shifted by one sample value and the next point is calculated. The correlation array is then normalised to consist of values between zero and one, by dividing each element by the total number of pulses in the standard Tx waveform, and the maximum with the associated time delay is found. These figures give a measure of the fit between the test and the standard device outputs, and the values obtained for the vowel shown in figure 1 are given above the appropriate Tx waveform plot.

From these figures, it would appear that the output from the peak-picker exhibits the 'best-fit' with the laryngograph output, the Gold-Rabiner the next best-fit, and the cepstral device the least best-fit. The delays associated with these measures indicate the time shift required to achieve that maximum correlation. These figures are presented to illustrate the development of the micro methodology, and they are not intended to give any more than an initial quantification between the devices. The nature of the speech input itself plays an important part in defining how appropriately a device will function, (1) and (2), and a suitable selection of speech data with which to test the devices must be gathered at a later date.

DISCUSSION

Fundamental frequency extraction devices operating in different domains from a speech input are designed to exploit various aspects of the input speech in their attempts to establish whether that input is voiced, voiceless or silent. It is, though, the very nature of the input speech waveform itself which thwarts the search for some universally applicable Fx measuring device. Even if one had any means available, however time consuming, there would still be the difficulty in deciding: exactly which portions of the speech waveform were voiced, voiceless or silent; the exact points at which to place boundaries between these portions; and exactly what the current measure should be for the fundamental period/frequency in a voiced segment at any given point in time. Thus there is no one measure which can be applied to assess the operation of a given device, such a comparison must be based on a carefully defined matrix of parameters chosen to quantify typical errors made by Fx estimation devices.

This study has been started with a view to eventually being able to establish such a parameter matrix. Analyses at the 'macro' level are already giving an overall measure of the a device's ability to estimate Fx, and both voiced and voiceless interval length. Whilst these measures are in themselves informative, they are not linked directly to the input speech, so for example, the modes in Vx plots obtained from the laryngograph and the cepstral device might each be made up of measurements made on different segments in the original input speech. Thus, whilst giving a useful initial device comparison, such global measures can not be used to determine the exact errors incurred with particular speech input sounds. It was with this in mind that the 'micro' measurements were begun.

Measurements at the micro level are intended to time align the device output waveforms before a comparison is made. The values given in figure 1 reflect, a least in part, the relative delay between the speech pressure and the Lx wavfeform at recording, due to the extra time taken for transmission in the acoustic path. In the case of the somewhat larger value gained from the cepstral process, this reflects the windowing used.

A correlation analysis is being used to achieve this, which essentially gives the devices the 'benefit of the doubt' in their ability to estimate Fx. This seems to be the most reliable parameter to use since most Fx estimation devices are primarily designed to do just that. A correlation based upon say, the length of estimated voiced or voiceless segments would not be as suitable, since there are occasions when the 'standard' output can sometimes have substantially longer or shorter values (see below) than the test outputs. The correlation value obtained can be used to give a quantitative measure, and it is interesting to note the highest figures are obtained from the time domain devices, where no windowing is involved, and the best of these is from the peak-picker where no output smoothing rules are employed (7, 12). The correlation values quoted in figure 1 might at first sight seem rather low, given that their possible range is from zero to one. This can be explained in terms of Fx jitter (1) in the outputs from the acoustically based devices, which is typically caused by noise and rapid formant transitions. Since in all Tx waveforms used there is just one non-zero value for each output pulse, and the sampling rate is 12.8kHz, quite small period measurement changes due to jitter could 'move' pulses, and therefore the resulting correlation will be lowered. Clearly this effect could be allowed for by making either the standard or the test pulses wider (two or three samples), but such a decision must wait until more experience has been gained.

All these measures depend on the supposition that an Lx measure is an appropriate standard measure. In practice, apart from the extremely few speakers for whom it is impossible to obtain a usable Lx output, the measure is highly reliable. However, for these comparison studies, it is the Tx waveform which is basic to their success, and during this study, a particular feature of this conversion is reckoned to be worthy of note. Figure 1 shows an Lx and speech pressure waveform along with Tx waveforms from various devices. The very first pulse in the Tx waveform derived from Lx, in this case using a Masscomp implementation, is separate from the rest. It occurs as a result of the precursive larynx adjustment prior to voicing, a feature which is shown on the Lx waveform in a manner similar to a typical closure-opening sequence in normal voicing. The figure illustrates that there is no acoustic effect resulting from this adjustment, and therefore none of the acoustically based devices will have an equivalent Tx pulse. Clearly these 'extra' pulses in the standard output will affect any statistical results which depend on the total number of Tx pulses. In an informal study using a voiscope, cases were found where more than one pulse was generated as a result of this feature, and this is currently under investigation.

In conjunction with this effect, the figure also shows that when voicing ends, for this speaker the amplitude of the last few Lx cycles is significantly lower than the others and that there is still a visible acoustic output. There are no Tx pulses from any device associated with these, so in this case the Lx to Tx would appear to be ideal for device comparison, but is it truly a standard? Similar cases have informally been observed where the amplitude of Lx drops to a level were its Tx conversion ceases, but the speech pressure waveform is such that acoustically based devices comtinue to produce Tx outputs. This effect is also under investigation, and in this case the standard Tx will have an inappropriately smaller total number of pulses.

Finally, it has been observed (12) that especially during plosives with a fully or partially voiced 'hold' phase, the Lx output is maintained whilst there is no output from acoustically based devices. In this case the standard Tx will have extra pulses. These effects will cause the standard Tx to bias the statistical calculations, for example the KS statistic shown in figure 4. Hence, Dx, Cx, Sx and Vx distributions (see figures 3, 5, 6) cannot reliably be used for device comparison until these problems with the standard Tx are cured.

CONCLUSIONS AND FUTURE WORK

The development of techniques designed to give, eventually, a quantitative assessment of the operation of fundamental frequency (Fx) estimation devices against a "standard" - the laryngograph (2) - has been described. Measures have been presented which are made at a "macro" (whole passage input) and a "micro" (single phone input) level, and typical results are given. It has been further shown that no single measure can be used to assess completely the operation of a given device.

In implementing these measures, examples have been isolated which illustrate that current techniques used to derive a fundamental period (Tx) measure from the laryngograph output waveform (Lx) require further investigation towards a more rigorous definition.

It is intended in the next stage of this work to utilise a best-fit estimate from the micro measure as the basis for time-aligning the Tx outputs with the standard before further processing. The exact nature of this processing has yet to be completely defined, but the macro measures provide a starting point since they quantify measurement categories already established (2). The results of such an analysis will be multi-dimensional, perhaps in matrix format. This reinforces the very problem being quantified, in that device optimisation is application specific (1, 2, 4, 7, 11, 12), and thus some parameters require extra attention in some cases but less in others. Thus it is felt that this is an appropriate course to be taking towards a comprehensive quantitative assessment of the operation of real-time speech fundamental frequency extractors.

ACKNOWLEDGEMENTS

The authors would like to thank: Jill House for making time to be recorded; Lynn Whitaker, John Mason, David Pearce, Andy Eaton and Neil Dove for work on the implementations of the digital devices; Peter Davies, Mark Huckvale, Mike Johnson and other SPAR colleagues for other software help. The SPAR component of this work is supported by SERC-Alvey grant-MMI/056.

REFERENCES

1. Hess, W., 1983, Pitch determination of speech signals, Springer-Verlag, Berlin.

2. Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. and McGonegal, C.A., 1976, IEEE Trans., ASSP-24,399-413.

3. Fourcin, A.J., and Abberton, E.R.M., 1971, Med. and Biol. Illust. 21, 172-182.

4. Hess, W. and Indefrey, H., 1984, Proc. ICASSP-84, 1-4.

5. Noll, A.M., 1967, J. Acoust. Soc. Amer., 41, 293-309.

6. Gold, B. and Rabiner, L.R., 1969, J. Acoust. Soc. Amer., 46, 442-448.

7. Howard, D.M., 1985, Ph.D. thesis, University of London.

8. Fourcin, A.J., Douek, E., Moore, B., Rosen, S.R., Walliker, J.R., Howard, D.M., Abberton, E.R.M., Frampton, S., 1983, An. New York Acad. Sci., 405, 280-294.

9. Fourcin, A.J., 1981, ASHA Reports 11, 116-127.

10. House, J., 1985, Improvements to speech synthesis-by-rule algorithms. Progress reports 1-3, research agreement No. F7P-50574-C.

11. Howard, D.M. and Seligman, P.M., 1983, Initial comparisons between two simple time domain fundamental frequency extractors. Speech, Hearing and Language: Work in Progress, 1, London: UCL, 97-105.

12. Howard, D.M. and Fourcin, A.J., 1983, Electronics Letters, 19, 19, 776-778.

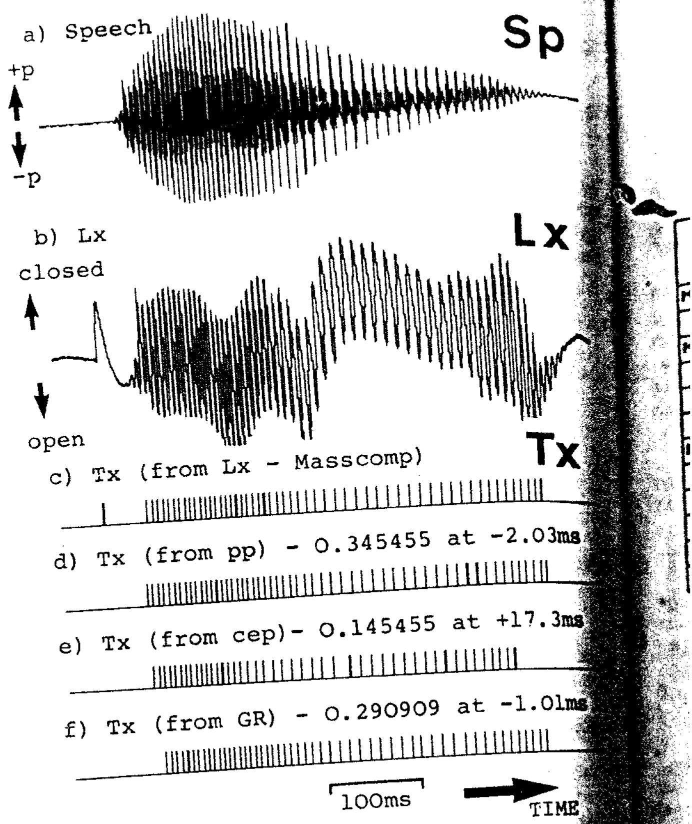

FIGURE 1: Tx waveform outputs from four devices, with speech and Lx - [

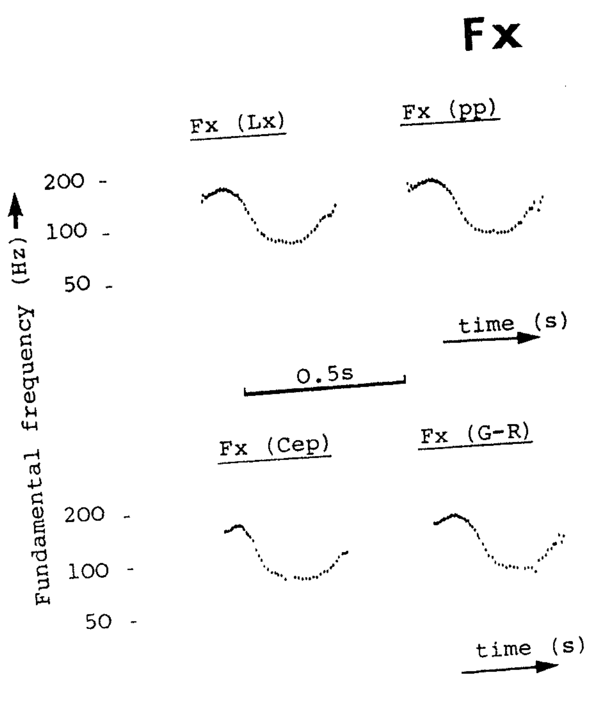

FIGURE 2: Fx contours from four devices. (plotted from Tx shown in fig 1)

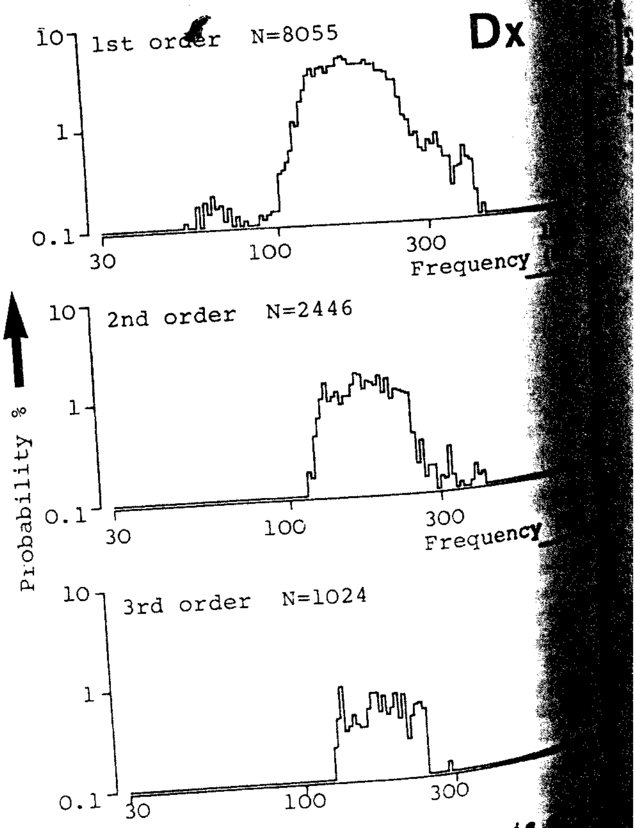

FIGURE 3a: Dx plots from Lx analysis passage read by a female sp

TYPICAL STATISTICS TABLE

FOR THE Dx PLOTS SHOWN ABOVE -

a) Female speaker
b) Read passage
c) Laryngograph output

Title:		ORDER	
	1	2	3
Mode (Hz)	167	167	134
Mean (Hz)	176	183	184
S.D. (log Hz)	0.158	0.112	0.108
	122/253	141/238	143/236
Median (Hz)	175	180	181
80% (Hz)	127-253	134-243	134-240
	(126)	(109)	(106)
90% (Hz)	111-313	129-303	130-290
	(202)	(174)	(160)
SAMPLE SIZE	8055	2446	1024

FIGURE 3b: Dx statistics.

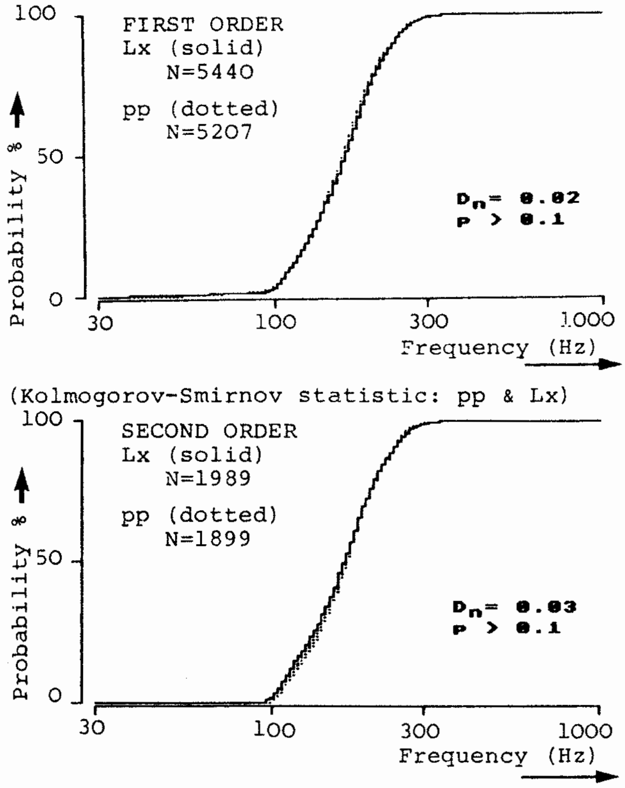

FIGURE 4: Cumulative Dx plot comparing peak-picker (pp) with laryngograph (Lx).

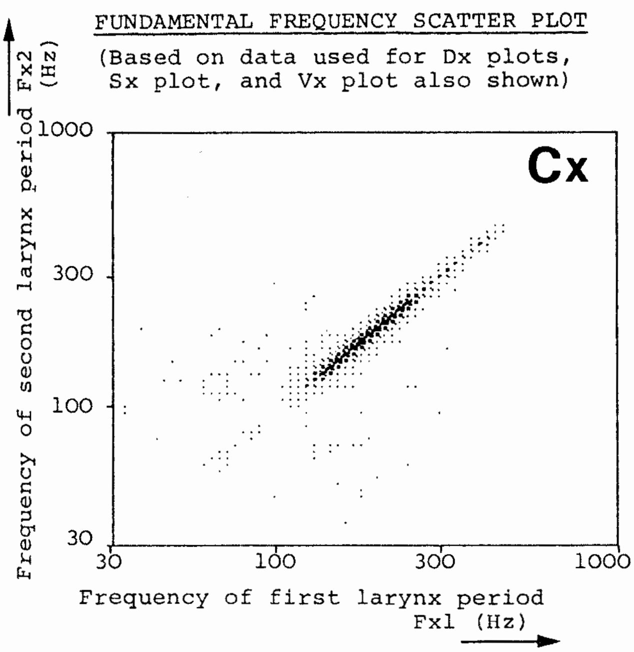

FIGURE 5: Larynx period scatter plot - Cx.

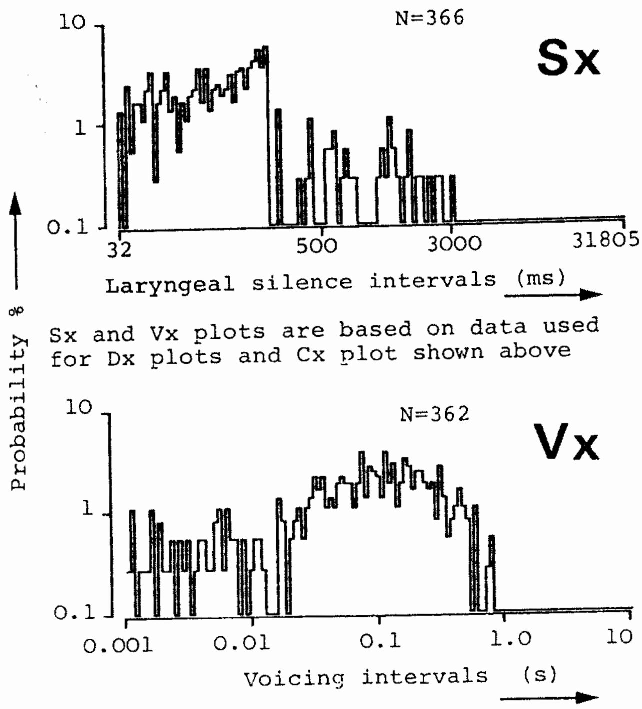

FIGURE 6: Laryngeal silence distribution - Sx & voicing interval distribution - Vx.