Proceedings of the Institute of Acoustics

A survey of fundamental frequency estimation techniques used in forensic phonetics

D.M. Howard[1*], A. Hirson[2*], J.P. French[3*] and J.E. Szymanski[4]

(1) Signal Processing: Voice and Hearing Research Group, Electronics Department, University of York, Heslington, York YO1 5DD, UK.
(2) Speech Acoustics Laboratory, Department of Clinical Communication Studies, City University, Northampton Square, London EC1V 0HB, UK.
(3) J.P. French Associates, Forensic Tape Laboratory, 156 Fulford Road, York YO1 4DA, UK.
(4) Applied Electrmagnetics Research Group, Electronics Department, University of York, Heslington, York YO1 5DD, UK.
(*) Founder Member of the International Association for Forensic Phonetics.

1. ABSTRACT

Forensic phonetic investigations are routinely concerned with the analysis of recorded speech which is telephone bandwidth limited. These analyses must always entail a detailed auditory-phonetic component undertaken by a trained and experienced phonetician. It is nowadays generally recognised that for forensic purposes, auditory-phonetic analyses must be complemented by acoustic measurements. One important aspect of speech transcription/analysis is the pitch pattern (qualitative) or fundamental frequency contour (quantitative). In the quantitative analysis of fundamental frequency, a number of devices/algorithms exist, each with relative strengths and weaknesses in terms of output reliability and accuracy. This paper reviews the basic operating principles of devices/algorithms found in forensic phonetics laboratories and illustrates their outputs for a high quality recording and the same speech via a standard telephone line.

2. INTRODUCTION

Comparison of tape recorded voices for forensic purposes usually requires matching segmental (phonemic), suprasegmental (e.g. intonation, tempo, rhythm) and extralinguistic features of speech. The latter includes voice characteristics, such as the overall fundamental frequency (f0) range of an individual, which are not utilised directly to code linguistic or phonetic contrasts. The measurement of f0 therefore has potential forensic significance with regard to suprasegmental features such as intonation and the speaker's long-term pitch range. In practice, these parameters may be important if they provide both high inter-speaker differences and low intra-speaker variance as twin criteria for speaker verification.

The usefulness of a speaker's measured f0 distribution has been shown to be: (a) a function of the duration of the sample used, and (b) context-sensitive. Barry et al. [1] have demonstrated that a minimum speech sample duration of 90 seconds is required in order to attain a stable f0 distribution. It is further shown that conversational speech may be highly variable in different situations and different from read speech, which is often provided as *reference* material in forensic casework. Speech contexts in which f0 distribution differences have been noted include the acoustic situation in which the speech was produced (e.g. [2]), the emotional state of the speaker (e.g. [3]), and the nature of the speech material, for example, police interviews frequently exhibit a lower mean and/or modal f0 value than that for a disputed telephone recording for the same speaker and this difference has been demonstrated experimentally [4]. Similarly, gender and age of the speaker [5] and tape speed at recording and playback [6] have relatively predictable effects on

speech f0. Since few of the above parameters are known for a disputed speech sample and therefore cannot be replicated for reference purposes, f0 data can only play a limited rôle in speaker identification.

Even when the speech material, speakers and the environment are controlled as far as possible, f0 measurements may still be difficult to replicate across forensic laboratories. This is a function firstly of the f0 estimation algorithms used, and secondly of their application and the interpretation of the results. This paper considers f0 estimation techniques commonly used in forensic casework and the results obtained for telephone and non-telephone speech data.

3. FUNDAMENTAL FREQUENCY ESTIMATION

The estimation of speech fundamental frequency (f0) has been the subject of research since the 1920's. (This field is often referred to as 'pitch extraction' in the literature, but since 'pitch' in this context is *subjective* and none of the algorithms to be discussed are modelling human pitch perception, the term 'f0 estimation' will be used in this paper.) No single device exists which estimates f0 accurately (as compared to a reference device) for any individual speaking in every acoustic environment. The choice of an f0 estimation technique should be made with direct reference to the application under consideration.

Techniques used for f0 estimation can be categorised into those which operate in: (a) the time domain, (b) the frequency domain, and (c) both time and frequency domains, referred to as 'hybrid' techniques. The input can be from a conventional microphone, a neck contact microphone, or from transducers which monitor larynx activity directly.

During speech, pitch is associated with sounds during which the vocal folds vibrate; the 'voiced' sounds, such as all the vowels and, for example, the consonants of *zoo*, *they*, *jay*, and *bee*. The human larynx cannot produce sounds on a perfect monotone; even a trained singer of early (pre-Renaissance) music singing a steady note produces some variation in f0. The f0 of voiced sounds is thus described as being 'quasi-periodic'. Time domain techniques exploit the fact that the waveform of a periodic sound is repetitive and they are designed to detect features which are usually found once per cycle during voiced sounds such as the major positive (or negative) peak, positive- (or negative-) going zero crossings, or the slope changes associated with major peaks.

The speech signal is often pre-processed by means of filtering or peak or centre clipping in order to eliminate most of the effects of the vocal tract resonances (formants) and the voiceless energy. Frequency domain techniques take advantage of the fact that a periodic signal has a harmonic amplitude/frequency spectrum. They make an estimate of f0 based on the pattern of spacing between the harmonics. Hybrid techniques make use of a combination of time and frequency domain features. Hess [7] gives a comprehensive review of f0 estimation techniques. Most of the errors associated with the various f0 estimation techniques are due to: (a) the *quasi*-periodicity of voice speech signals, and/or (b) the difficulties in detecting the transitions between voiced to voiceless and voiceless to voiced segments.

The electrolaryngograph [8] is a well-established reference f0 estimation device (e.g. [9, 10]). It measures the changing high frequency electrical impedance between two electrodes place superficially on the neck at the level of the larynx during voiced speech. It is thus immune to the effects of external acoustic noise. It should be noted that there are a few subjects for whom it can be extremely difficult to obtain a usable electrolaryngograph output (Lx) waveform, due for example, to excess neck fat tissue, but for the vast majority of subjects a reliable Lx waveform can be obtained. This paper compares f0 contours produced by six algorithms commonly used by forensic phoneticians with Lx-based reference f0 data

A SURVEY OF F0 ESTIMATION TECHNIQUES USED IN FORENSIC PHONETICS

4. EXPERIMENT

For the purposes of this paper, two recordings of a small portion of a read passage were used. The intention was not to attempt to characterise the f0 for the speaker (in which case a minimum speaking time of 90 seconds would be appropriate), but rather to compare the performance of different algorithms with identical speech material. Thus any length of speech material could be used, and the short section *whenever his friends asked him if he would like to go out with them* was used. The speaker is an adult male with no known speech of hearing problems. The two recordings were made synchronously, one at each end of a telephone line. At the speaker's end, denoted in this paper as 'transmit', a stereo digital audio tape (DAT) recording was made of the output of a Sennheiser ME40 microphone on one channel, and the Lx waveform on the other. At the listener's end, a mono DAT recording was made by means of a telephone monitoring device, denoted in this paper as 'telephone'.

The devices used were as follows:

(TSA) a time domain 'temporal structure analysis' system (*Signalyze* Mac-based package), which identifies strong falls and rises characterising the onset and offset of large oscillations in voiced speech;
(FFT) a frequency domain fast Fourier transform (FFT) based technique (*Signalyze* Mac-based package), which measures the frequencies of harmonics;
(AUTO) a time domain technique autocorrelation technique (*Signalyze* Mac-based package), which correlates the amplitude values of a given stretch of signal with those of succeeding segments;
(CEP) a frequency domain cepstrum technique (*LSI Speech Workstation* PC-based package), in which a spectrum of the log power spectrum of the speech signal is calculated [11];
(PP) a time domain peak-picker [12] available in hardware form as part of the electrolaryngograph system, which locates major peaks in the speech pressure waveform;
(MSL) a time domain technique (*Micro speech lab* PC-based package), which looks for dominant quasi-periodic activity in terms of the waveform slopes around the peaks; and
(Lx) the electrolaryngograph ('transmit' end only).

It should be noted that in the cases other than **Lx**, **PP** and **CEP**, the algorithms are as described in the relevant 'User Manuals'. Unfortunately, no further references are provided.

5. DATA ANALYSIS AND RESULTS

All f0 estimation algorithms have to tune-in to the amplitude of the input speech data. For those which operate in non-real-time, this process can be automatic, but may necessitate an additional data pass. For most, the input level (threshold) may be manipulated by the user. For most time domain devices, the threshold setting can have a highly significant effect on the f0 output, particularly when the speech data has noise associated with it. In this experiment, the data recorded at the telephone end incorporates a degree of line and other noise. Figure 1 shows the speech pressure waveform for the utterance with the f0 contours obtained from the autocorrelation (AUTO) algorithm available in the Mac-based *Signalyze* system for different threshold value settings (20%, 25%, 30%, 35% and 40% of the maximum to minimum signal amplitude).

In order to appreciate the details of the f0 contours shown, reference should be made to the **Lx** coutour plotted in figure 2 which is our reference contour. If the **Lx** output is compared with the **AUTO** output for a threshold setting of 40% (lowest contour in figure 1), it can be seen that four segments of the contour are conspicuous by their absence: the fall following the initial f0 rise, two adjacent segments around the centre of the contour, and the last segment. At all other threshold values, one of the centre segments is output, but it is higher in frequency than the reference. At

values of 20%, 25% and 30% the fall following the initial rise is output, and the other centre contour is output, but it too is higher than the reference. The final segment is only apparent for threshold values of 20% and 25%, but it is higher than the reference and it is shorter in duration. The f0 contours for threshold values of 20% and 25% exhibit considerable noise which is mainly well above the contour itself. It is thus vital to view contours when carrying out f0 analysis and to have a basis on which to check their reliability. In general, of course, there is no reference available, and this check has to be based on critical listening to the speech by a trained phonetician.

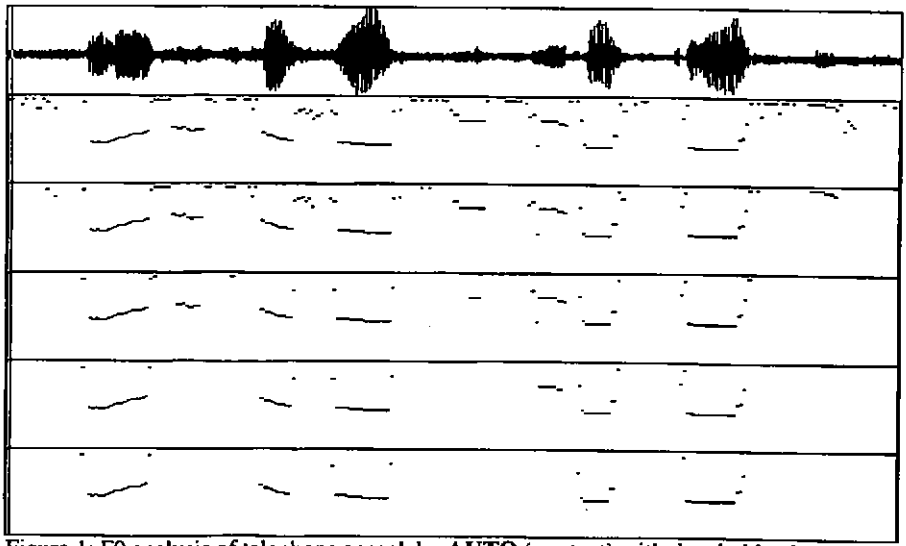

Figure 1: F0 analysis of telephone speech by **AUTO** (see text) with threshold values of 20%, 25%, 30%, 35%, and 40%. (Speech waveform at the top.)
{NOTE: Output f0 smoothing is disabled. X axis: time; Y axis: f0.}

In summary, as f0 analysis is carried out for speeech degraded by the addition of noise and/or reverberation, or transmission via a telephone line, the threshold setting (automatic or manual) becomes a vital parameter to set appropriately with due regard for the potential f0 errors which can be incurred. In particular, there will be a spurious f0 output given for voiceless sounds and acoustic noise when the threshold value is set too low, and voiced segments tracked by Lx may be missing in the f0 contour and segments will tend to be clipped in time when it is set too high. Time domain devices are particularly susceptible to f0 doubling errors (i.e. f0 estimations an octave higher than may be expected), usually due to a prominent second harmonic associated with sounds with low first formants. Frequency domain devices, on the other hand, are often not as accurate on a micro f0 level since the accuracy of cycle-by-cycle changes in f0 is compromised by the requirement to window the speech data to enable conversion to the frequency domain.

Many of the f0 estimation packages available incorporate output f0 contour smoothing designed to prevent jumps in f0, particularly isolated octave errors. In the case of the **AUTO** algorithm threshold changes illustrated in figure 1, output smoothing is available but here it was disabled to illustrate some of the f0 errors. Output smoothing should be used with care, since it can also serve to distort the f0 data, for example, during creaky or hoarse voice or abnormal modes of phonation

A SURVEY OF F0 ESTIMATION TECHNIQUES USED IN FORENSIC PHONETICS

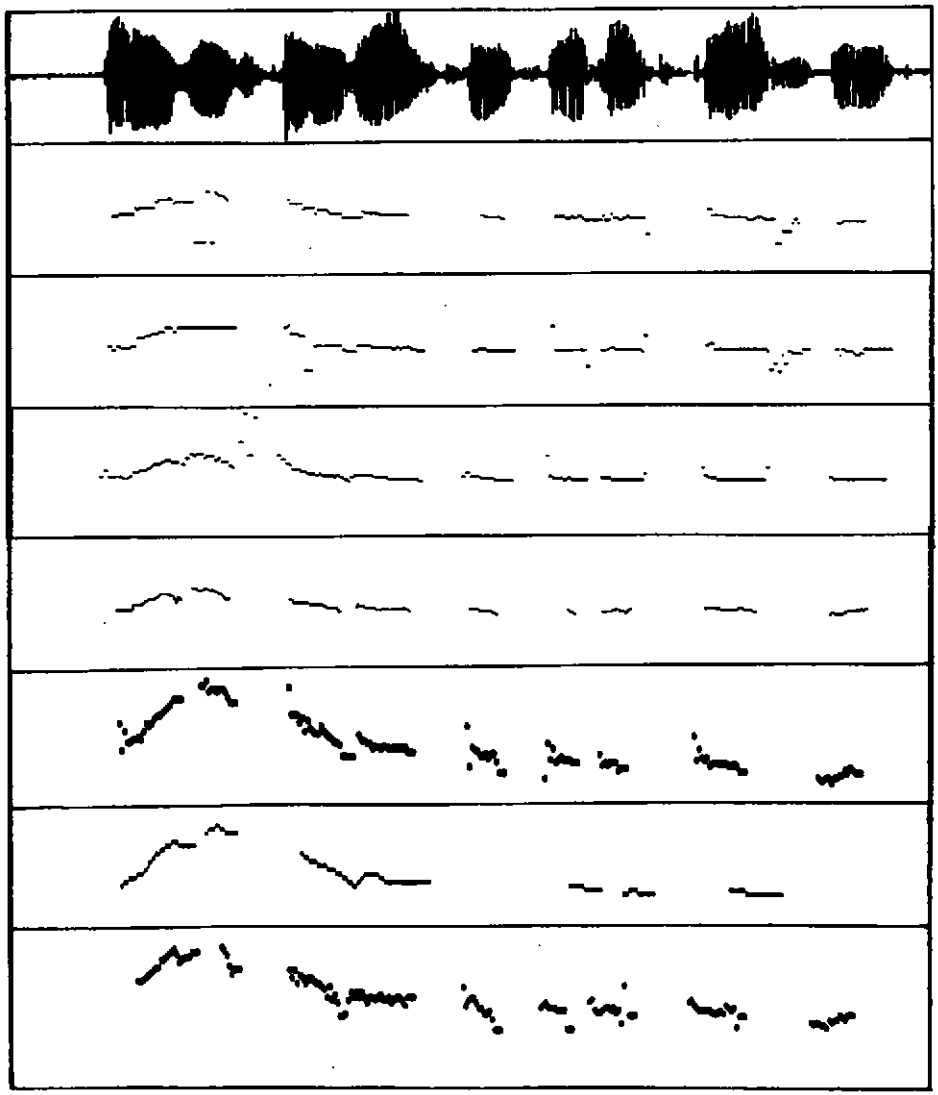

Figure 2: Speech pressure waveform (transmit end) with f0 contours from: **TSA**, **FFT**, **AUTO**, **CEP**, **PP**, **MSL**, and **Lx** (reference contour).
NOTE: Frequency scales are NOT similar, time scales are approximately lined up.

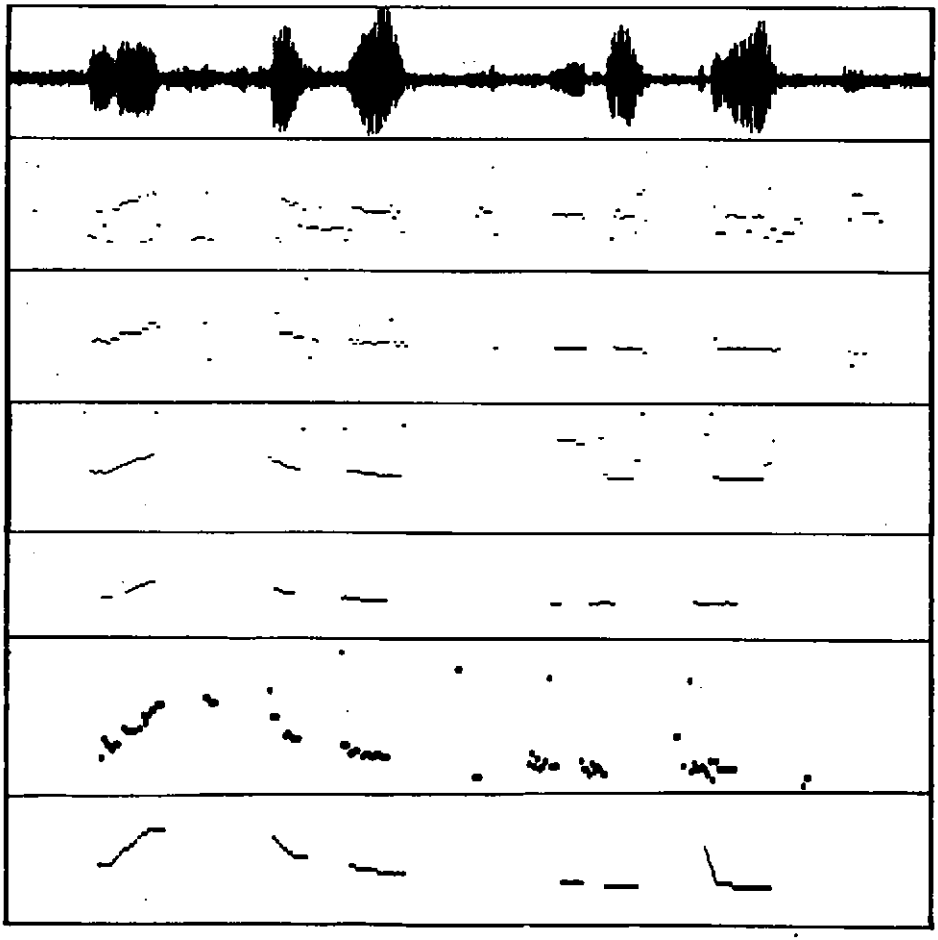

Figure 3: Speech pressure waveform (telephone end) with f0 contours from: **TSA**, **FFT**, **AUTO**, **CEP**, **PP**, and **MSL**.
NOTE: Frequency scales are NOT similar, time scales are approximately lined up.

associated with voice pathologies. Extreme values of output smoothing can distort quite normal f0 changes, and while the output contour will look smoother (hence the name), it may be less representative of the f0 contour than the non-smoothed version.

Figure 2 shows the f0 contours produced by all the devices as well as the Lx output reference contour for the speech recorded at the transmit end, and figure 3 shows the outputs from all devices based on the speech recorded at the telephone end. (It should be noted that only the first three contours in each figure can be directly compared as they are all plotted from one Mac-based package. The other contours have been plotted by a variety of software packages and the frequency scales are NOT equivalent, since the plots have been enlarged/reduced to make the time scales approximately equivalent.) In each case the threshold level has been arrived at heuristically, adjusting the level until it is deemed to produce minimum errors in the f0 output.

Comparison of the reference f0 contour with that obtained for the acoustically-based devices for the transmit end (see figure 2) shows that all devices except **MSL** have located all the f0 segments in the utterance, and that the f0 patterns generally agree. The **MSL** device gives an output in which two of the f0 segments are missing. More detailed examination of the contours for each device reveals that in some cases, particularly **TSA** and **CEP**, some of the segments are shorter in time than the reference. In the case of **CEP** this is most likely due to the f0 smoothing rules (not available to the user) applied at the onset of voicing. The **Lx** and the **PP** outputs are based on a cycle-by-cycle analysis of the speech data which can be observed as a 'rougher' contour for the transmit end. (It should, however, be noted that they have the most magnified frequency scale due to the time scaling during figure preparation.) The outputs from **TSA**, **FFT**, and **AUTO** show evidence of f0 errors where there are values which are scattered away from the main contour. The output from **FFT** exhibits sections of constant f0, especially immediately following the initial rise. These are inappropriate representations of f0 for human speech, but the **FFT** algorithm has benefits when the input speech is noisy (see below).

Device	Transmit end	Telephone end
	mean(Hz) {SD(Hz)}	mean(Hz) {SD(Hz)}
Lx	137.2 {16.7}	n/a {n/a}
TSA	136.0 {18.5}	126.5 {26.8}
FFT	139.0 {18.5}	145.4 {25.9}
AUTO	147.8 {20.5}	162.7 {36.3}
CEP	139.1 {16.2}	139.1 {13.7}
PP	141.0 {20.8}	137.2 {13.2}
MSL	143.0 {28.0}	140.0 {29.0}

Table 1: F0 statistics from six devices from both ends of the telephone as well as reference Lx data from the 'transmit' end.

Comparison of the f0 contours for the telephone end (see figure 3) shows that all devices miss some of the f0 segments and that there are considerable spurious f0 values introduced into the contour, particularly for **TSA** and **AUTO**. Devices which produce the 'cleanest' looking f0 contours are **FFT**, **CEP** and **MSL**, but they also have f0 segments which appear to be artificially constant, and often too short when compared to the reference. It is interesting to note that the f0 segment near the centre and that at the end of the utterance are almost lost in the noise (see the speech pressure waveform in figure 3 and compare with that for figure 2). The only devices which

give any output for either of these are **TSA**, **FFT** and **PP**. On a visual representation of the f0 contours, the **FFT** technique appears to give the most appropriate f0 contour for the telephone end, but it provides a rather poor representation for the transmit end (see above).

The mean f0 values and the standard deviations are given in table 1 and plotted in figure 4 for all devices. Although these means and standard deviations are based on a speech utterance which is far too brief to characterise the f0 for a speaker, they can be directly compared since each is based on an analysis of the same speech data. The **PP** values given are based on a second order analysis [13] which accepts adjacent f0 values only if they fall in the same histogram bin, thus providing a degree of f0 smoothing.

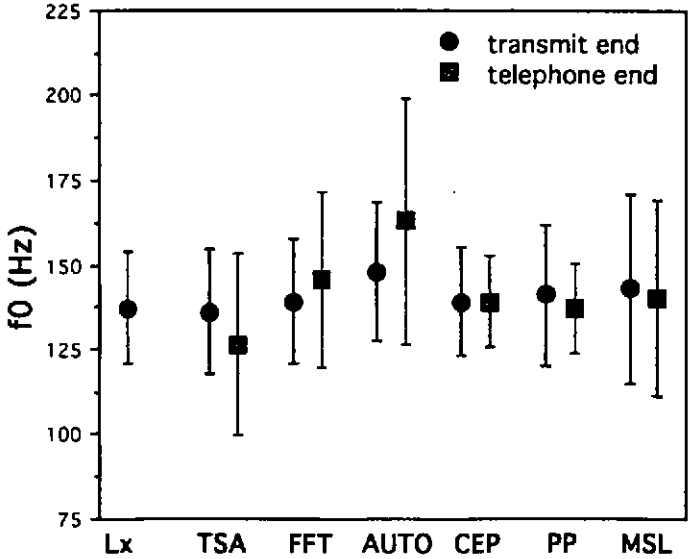

Figure 4: Graph of f0 means and standard deviation error bars for all devices.

Since each algorithm is designed to provide information about f0 of the input speech waveform, it is possible to hypothesise that the six f0 mean values form a sample from a large population of f0 means from similar/related algorithms designed to estimate f0. Assuming these six means to be consistent with a normal population with the reference **Lx** mean of 137.2Hz, it is possible to estimate the population variance to be 4.43 and to apply a t-test to confirm that this is a reasonable assumption at the 5% significance level (the 95% confidence limits are 140.55Hz±4.65Hz). Similarly the six telephone speech f0 means are also consistent with a normal distribution of mean 137.2Hz (95% confidence limits of 141.82Hz±12.56Hz).

In *both* cases, the f0 means from **AUTO** lie *outside* the confidence limits and are also borderline even at the 1% significance level. This indicates that the performance of **AUTO** is notably inconsistent with the others. The f0 means from the subset of algorithms excluding **AUTO**, leads to 95% confidence limits of 139.10Hz±3.68Hz for the transmit end, and 137.64Hz±8.61Hz for the telephone end. It can then be seen that the **TSA** algorithm is the next most extreme in its f0 estimates.

A SURVEY OF F0 ESTIMATION TECHNIQUES USED IN FORENSIC PHONETICS

6. CONCLUSIONS

A number of fundamental frequency estimation techniques have been introduced and applied to a speech utterance recorded simultaneously at each end of a telephone line. These techniques are commonly found in forensic phonetics laboratories running on personal computers, and the analysis of telephone speech is a common component of forensic phonetics casework. It has been shown that the output from f0 estimation devices is highly variable between algorithms with errors appearing from: (a) omitted sections of f0 contours, (b) shortened f0 segments, (c) lengthened f0 segments, (d) f0 segments with spurious values scattered away from the true value, (e) segments with constant f0 which are unlikely to be found in human speech, and (f) inappropriately set threshold (and other) user parameters. These findings are in keeping with the notion that the choice of f0 estimation techniques for a given situation should be made with reference to the situation itself and the errors which are acceptable/non-acceptable. Simple inspection of the f0 data derived from short speech utterances is inadequate for comparative purposes. Errors may be introduced by the performance of f0 estimation devices, due to their specific mode of operation and/or the particular settings employed by the user.

7. ACKNOWLEDGEMENTS

The authors would like to thank David Rossiter who took part in this experiment. This work is part supported by the International Association for Forensic Phonetics.

8. REFERENCES

[1] Barry, W.J., Goldsmith, M., Fourcin, A.J., and Fuller, H. (1990). 'Larynx analyses on normative reference data',

[2] Summers, W.Van., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., and Stokes, M.A. (1988). 'Effects of noise on speech production: acoustic and perceptual analysis', *Journal of the Acoustical Society of America*, **84**, (3), 917-928.

[3] Williams, C.E., and Stevens, K.N. (1972). Emotions and speech: Some acoustical correlates, *Journal of the Acoustical Society of America*, **52**, (4), 2, 1238-1250.

[4] Hirson, A., French, J.P., and Howard, D.M. (1993 - In press). 'Forensic aspects of telephone speech: preliminary findings', In: Windsor Lewis, J., (Ed.), *Studies in general and English phonetics: Essays in honour of Professor JD O'Connor*, London: Routledge and Kegan Paul.

[5] Aronson, A.E. (1985). *Clinical voice disorders: An interdisciplinary approach*, 2nd Ed., New York: Thieme Inc.

[6] Hollien, H. (1990). *The acoustics of crime. The new science of forensic phonetics*, New York: Plenum Press.

[7] Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices*, Berlin: Springer.

[8] Fourcin, A.J., and Abberton, E.R.M. (1971). *First applications of a new laryngograph*, Medical and Biological Illustration, **21**, (3), 172-182.

[9] Hess, W., and Indefrey, H. (1984). Accurate pitch determination of speech signals by means of a laryngograph, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, **ICASSP-84**, 1-4.

[10] Howard, D.M. (1989). 'Peak-picking fundamental period estimation for hearing prostheses', *Journal of the Acoustical Society of America*, **86**, (3), 902-910.

[11] Noll, A.M., (1967). Cepstrum pitch determination, *Journal of the Acoustical Society of America*, **41**, 293-309.

[12] Howard, D.M., and Fourcin, A.J. (1983). 'Instantaneous voice period measurement for cochlear stimulation', *Electronics Letters*, **19**, (19), 776-779.

[13] Abberton, E.R.M., Howard, D.M., and Fourcin, A.J. (1989). Laryngographic assessment of normal voice: A tutorial, *Clinical Linguistics and Phonetics*, **3**, (3), 281-296.