PITCH ANALYSIS METHODS FOR CROSS-SPEAKER COMPARISON

J A Maidment[*]
M L García Lecumberri[†]

[*]Department of Phonetics & Linguistics, University College London, UK
[†]Department of English, UPV-EHU, Vitoria, Spain

ABSTRACT

A system of fundamental frequency analysis and normalisation is described for obtaining pitch data and comparing them across speakers. This system was used for the analysis of English and Spanish speakers' productions in order to compare the realization of accentual focus in the two languages.

The system is based on the simultaneous recording of speech and the laryngeal signal. The latter is monitored by means of an electrolaryngograph. The analysis is done in three stages: (a) auditory analysis, (b) PCLX analysis obtaining measures in Hz for peaks, troughs and other relevant points in the contour and also fundamental frequency statistics from the complete data set for each speaker, (c) a detailed analysis using SFS with simultaneous display of speech pressure waveform, Lx waveform, excitation period measurements and fundamental frequency trace and with playback facilities for speech and Fo. The analysis described at (c) is used for segmentation, to filter out nodes which are due to micro-intonation, and to pinpoint problem areas in the Fo trace. Outlying and anomalous period measurements may be replaced by a five-point median value. The resulting contours are normalised by converting Hz measures to percentage values (positive or negative) of the speaker's mean Fo which is obtained from the analysis described at (b).

1. INTRODUCTION

Many studies of pitch variation in speech use methods which rely on the extraction of fundamental frequency from the speech waveform. Examples are cepstral analysis and waveform peak-picking algorithms. The method described here is based on the use of the electrolaryngograph which monitors laryngeal activity directly. Briefly, the laryngograph measures current flow between two electrodes placed externally on the speaker's neck at the level of the larynx. This flow is at a maximum when there is good contact between the edges of the vocal folds and is at a minimum when there is no contact between the folds. The output of the laryngograph, known as Lx, is a time-varying quasi-periodic signal whose amplitude at a given point is an analogue of the conductance of the speaker's neck and therefore of the degree of vocal fold contact at that instant. Further information on the laryngograph may be found in Abberton et. al. (1989). Lx and the speech waveform may be recorded synchronously for further processing. Two further representations of laryngeal activity which may be produced by such processing are (a) measures of the period of the Lx waveform, known as Tx and (b) measures of the instaneous fundamental frequency of excitation derived from Tx, known as Fx. The method described here uses Tx and Fx representations as its starting point and proceeds in three stages (i) auditory analysis (ii) a preliminary instrumental analysis using a suite of software called PCLX (iii) a more detailed instrumental analysis using software of the Speech Filing System (SFS). The rationale for this three part analysis will be dealt with as each component is described below.

2. AUDITORY ANALYSIS

The purpose of the initial auditory analysis is to obtain a record of the broad characteristics of the pitch movements in the speech sample. Repeated listening to short segments of the speech signal enables the sketching of impressionistic pitch contours as in the example in

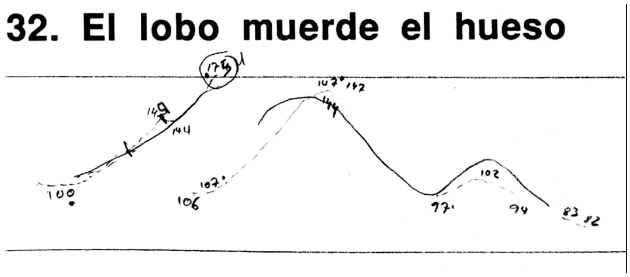

Figure 1 below. [IMAGE A141G01.GIF]

Figure 1: Auditory analysis of the sentence *El lobo muerde el hueso.* (=The wolf bites the bone) The numbers in Hz are derived from subsequent analyses. The dotted lines represent corrections to the original analysis.

3. PCLX ANALYSIS

PCLX (Laryngograph Ltd) is a suite of programs which runs on a PC. The initial processing entails the conversion of Lx to Tx. Files of Tx values for utterances may be stored and may be displayed as Fx contours against time. (See Figure 2). [IMAGE A141G02.GIF] There are also facilities for measuring Fx at any given point in the contour. The display facility was used to correct any major errors in the auditory analysis, bearing in mind that Fx contours contain variation due to micro-intonational effects. Such effects and other areas of doubt such as creakiness and jitter were noted for further detailed analysis. The Fx measurement facility was used to produce a set of node values in Hz for local maxima and minima in the contours.

PCLX was also used to produce global statistics for a speaker's voice.

the latter is an indication that the recording is likely to present major problems for further analysis.

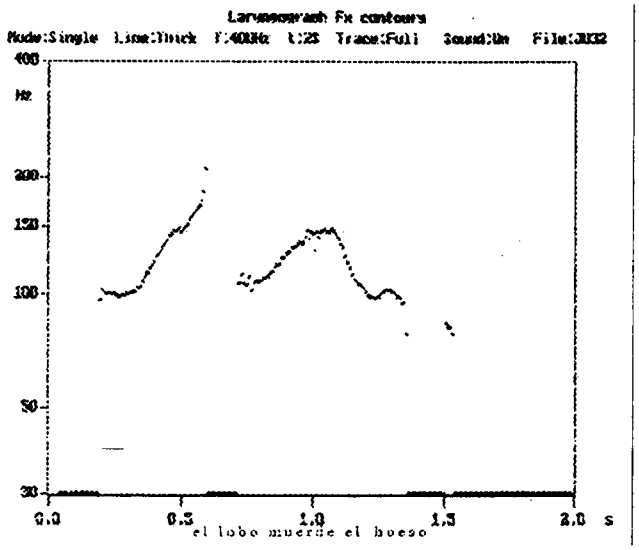

Figure 2: PCLX Fx contour of the utterance *El lobo muerde el hueso.*

To sum up, the results of this part of the analysis are a set of preliminary node values in Hz for each sample, a set of problem areas for futher investigation and a measure of overall mean fundamental frequency of each speaker's voice. It is also a simple matter to record internode durations at this stage.

4. SFS ANALYSIS

The Speech Filing System (SFS), developed at UCL allows the display, annotation, playback and measurement of speech signals and analyses

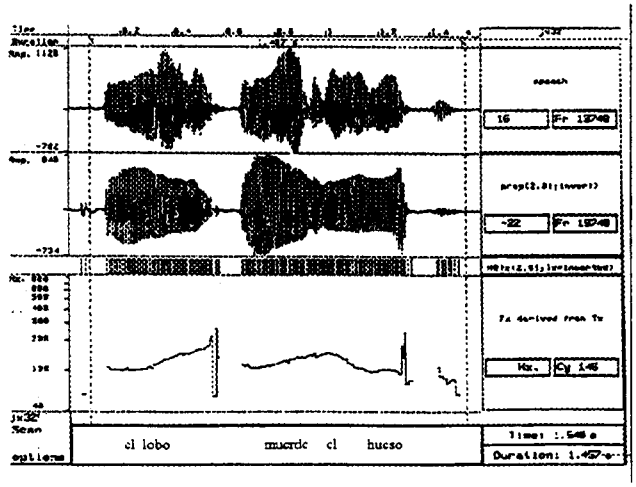

Figure 3: SFS speech, Lx, Tx and Fx display of the sentence *El lobo muerde el hueso.*

of these. More information about the system is currently available on the World Wide Web (http://www.phon.ucl.ac.uk/resources/sfs.html)

For the present purposes, four synchronised displays were used: speech pressure waveform, Lx waveform, Fx contour and Tx measurements. (See Figure 3). [IMAGE A141G03.GIF] Using the facilities SFS provides for zooming in to view portions of the signal at high magnification, the reliability of the Tx measurements could be verified by comparing them with the display of the laryngeal waveform recorded via the laryngograph. (See Figure 4). [IMAGE A141G04.GIF]

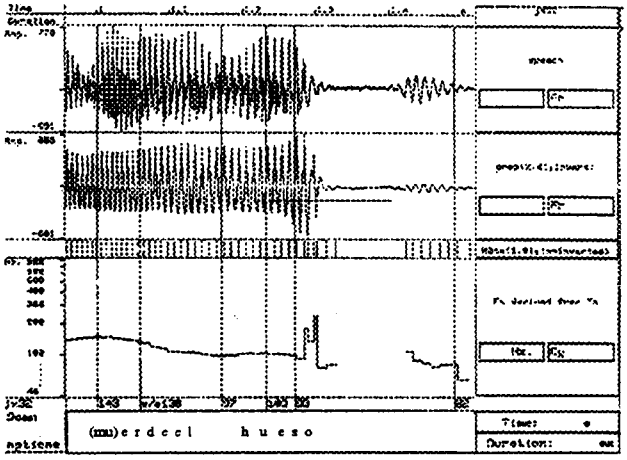

Figure 4: High magnification display of SFS Speech, Lx, Tx and Fx of the sentence *(El lobo) muerde el hueso.*

Any anomalies were discarded and replaced by the median of a five point window surrounding the anomalous measurement. SFS was also used to check the segmentation of the signal. The peak node values obtained by the PCLX analysis were examined to ensure that they fell within the boundaries of an accented vowel and trough nodes were checked to ascertain that they corresponded to sonorant consonant articulations. If a trough value was seen to fall within a voiced obstruent, it was discarded and a new value selected immediately before or after the Fx dip caused by the obstruent. Peaks following voiceless obstruents were also examined. It is well known that voiceless obstruent articulations have the effect of increasing Fx in the initial portion of a following vowel. See for example Lehiste & Peterson (1961) and Haggard et. al. (1969). The peak nodes in this environment were discarded and replaced by either a value from the earliest portion of the vowel where Fx was steady or, in cases where no such steady Fx could be discerned, the median of a five-point window containing the highest Fx values.

5. NORMALIZATION

The Hz node values for Fx obtained by the above procedures were normalized to the individual speaker's mean Fx which was obtained as part of the PCLX analysis. Each Hz value was converted to a

percentage of mean Fx using the simple conversion:

$$Fx' = 100*(Fx - mean)/mean)$$

This normalisation method is somewhat similar to that used by Jassem and Kudela-Dobrogowska (1980) and by Kelm (1987). This method has the advantage that it is extremely easy to compute and is readily understandable. As the perception of frequency is approximately linear in the range covered by the fundamental frequency of most speakers there seems little to be gained from a more complex normalisation method.

6. APPLICATION

The methods described above have been used in a study of intonational focus in English and Spanish. Full details may be found in García Lecumberri (1995). A brief account will be given here, concentrating on the analysis of utterance by Spanish subjects.

Four subjects, three male and one female, were selected out of an initial pool of 16 speakers who were recorded and whose recordings were subjected to the type of auditory analysis outlined in section 2 above. The speakers' accent was homogenous, being the variety of Northern Castilian Spanish spoken in and around Vitoria.

The material recorded consisted of replies which were read by the subjects in response to a stimulus sentence. These stimulus sentences were designed to elicit focus on various parts of the response sentences: broad focus, narrow focus on the verb phrase, focus on subject phrase and the like. For example, the stimulus sentence for the utterance in Figures 1 -4 was *¿Qué hace el lobo con el hueso?* (=What does the wolf do with the bone?), eliciting narrow focus on the verb. The sentences also had differing phonetic and syllabic structures, so that differences in pre-focal and post-focal intonational structures could be investigated

Some of the main findings of the study, which relied to a great extent on the analysis and comaprison methods described above, were:

1. Subject phrase focus was most consistently marked with an accent realised as a fall from high to low. The mean peak height for such falls was speaker's mean Fx + 18.5% and the mean value for the end of the fall was speaker's mean Fx - 15%.

2. Verb focus was consistently signalled by a double accent structure. The first accent on the subject phrase was realised as a rise from a mean value of speaker's mean Fx - 5.5% to a mean value of mean Fx + 27.8%. The accent on the focussed verb was realised as a fall with a mean peak value of mean Fx + 9.5% and a mean endpoint of mean Fx - 15.7%.

3. In contradistinction to English where post-focal accents are deleted, Spanish post-focal accents may remain, but show a reduced Fx obtrusion.

7. REFERENCES

1. Abberton, E.R.M., Howard, D.M. and Fourcin A.J. "Laryngographic assessment of voice quality: a tutorial." *Clinical Linguistics and Phonetics*3: 281-296, 1989.

2. García Lecumberri, M.L. *Intonational signalling of information structure in English and Spanish: A comparative study*. Unpublished PhD thesis. University of London, 1995.

3. Haggard, M., Ambler,S. and Callow, M. "Pitch as a voicing cue." *Journal of the Acoustical Society of America*, 47: 613-617, 1969.

4. Jassem, W. & Kudela-Dobrogowska, K. "Speaker independent intonation curves." *The Melody of Language: Intonation and Prosody*. Waugh, L.R. and van Schooneveld, C.H. (eds). Baltimore: University Park Press. 135-148. 1980

5. Kelm, O.R. "An acoustic study on the differences of contrastive emphasis between native and non-native Spanish speakers.*Hispania* 70: 627-633, 1987.

6. Lehiste, I. and Peterson, G.E. "Some basic considerations in the analysis of intonation." *Journal of the Acoustical Society of America*, 33: 419-425, 1961