RECENT RESEARCH ON METHODS FOR AUTOMATIC ESTIMATION OF VOCAL EXCITATION

J. S. GILL

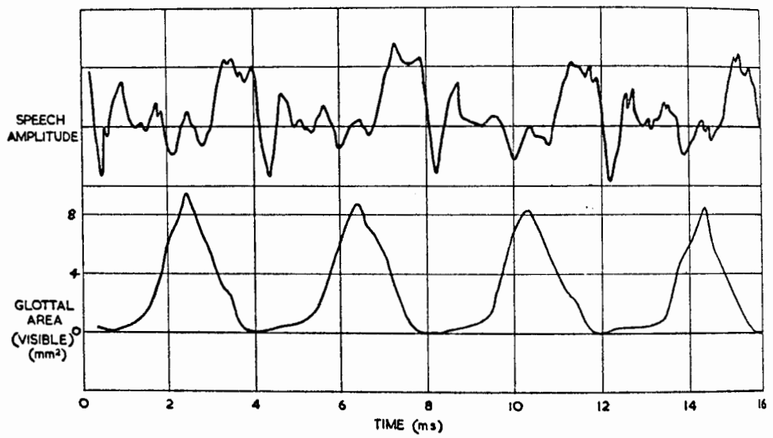
There are many applications, both in the study of speech and in the development of analysis-synthesis systems of telephony, for which a reliable automatic method of estimating the fundamental frequency of the vocal fold vibrations is required. In the past various methods of estimation have been developed but no wholly satisfactory solution has yet been found. The present paper describes some recent research in this field.

High-speed cine-photography has been used to study the variation with time of the area between the vocal folds (the glottis) and to relate this variation to the resulting sound pressure waveform. Fig. 1 shows some of the results which were obtained from a male subject when uttering a sustained /α/. The glottal area wave is, in this case, approximately triangular and periodic. Although small differences between the waveforms of successive cycles have been observed throughout this work it appears that the fundamental frequency of the glottal wave usually varies smoothly with time. The glottal variations modulate the air flowing between the lungs and the pharynx and the waveform of the modulated air stream is then shaped by the oral and/or nasal cavities of the talker to produce speech. The presence of the intervening vocal tract can be seen, on Fig. 1, to increase the difficulties of estimating the fundamental frequency.

The structure of speech and some of the difficulties of fundamental frequency estimation are revealed by short-term spectrum analysis. Two analyses of the phrase "there are many call(s)" are shown on Fig. 2. These spectrograms show both the harmonic structure of the glottal modulation and the manner in which the glottal flow is modified by the vocal tract. It can be seen that the acoustical properties of the vocal tract can vary rapidly, as for example in the transitions to and from the nasal consonants /m/ and /n/. Also it appears that, quite apart from the effects of the vocal resonances and anti-resonances, the envelope of the glottal spectrum does not vary smoothly with frequency. This phenomenon, which arises from the approximately triangular waveform of the glottal excitation, is evident on both of the spectrograms. In the case of the wide-band analysis the effect is revealed by the presence, within each glottal cycle, of two or three vertical striations.

The effects of the vocal tract are least evident in the band below 300 c/s. Consequently this region of the speech spectrum usually provides the most reliable indication

WAVEFORMS OF GLOTTAL AREA AND CORRESPONDING SOUND



FROM FILM B.R.S. 2.
SPEED ≃ 5000 FRAMES / SEC.
VISIBILITY 70% (APPROX.)

SPEECH TRANSDUCED BY CONDENSER MICROPHONE SITUATED APPROX. 4" FROM LIPS.

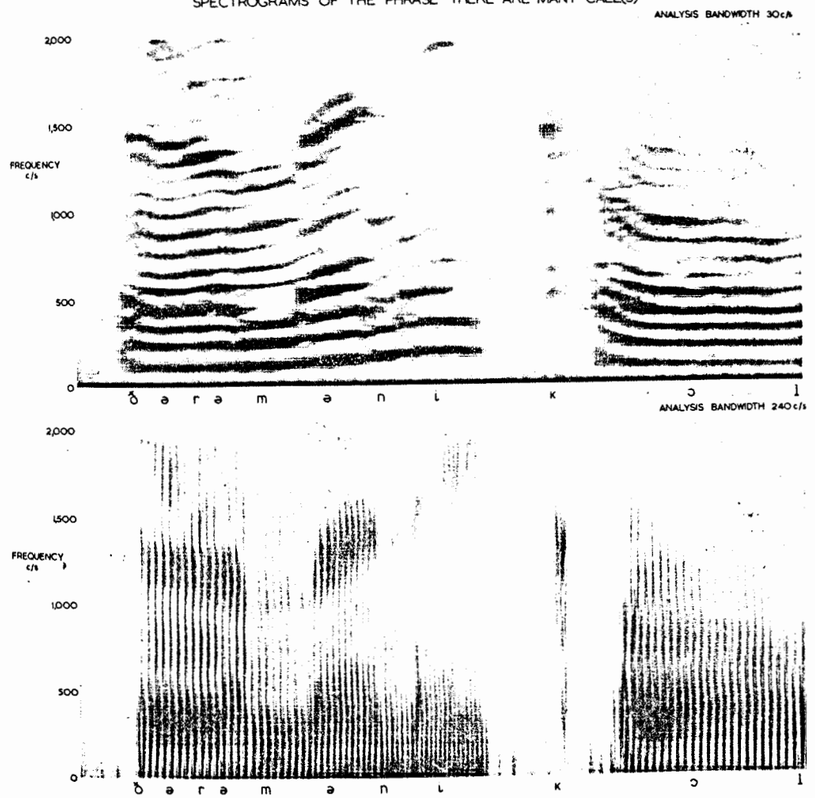
Fig. 1.

of the fundamental glottal frequency, and should therefore be used for this purpose whenever available. These low-frequency signals are, however, substantially absent from telephone-quality speech.

The effect of removing the band of frequencies below 300 c/s from the waveform of /mən/ is shown on the two upper traces of Fig. 3. An approximate estimate of the fundamental glottal frequency can be derived from the timing of the peaks of the waveform of the 0-3000 c/s band. When, however, the band below 300 c/s is removed the peak detection method fails almost completely. Even when the band 0-300 c/s is available the estimate provided by this method is only an approximation because the relative timing between the glottal wave and the speech varies with the changing configuration of the intervening vocal tract.

It has been suggested that, even when the fundamental glottal frequency of the speech is absent, a component at this frequency can be regenerated by means of intermodulation. A study has been made of two methods based on full-wave rectification of the available speech band and on the detection of the envelope of the speech. The envelope is derived by first translating the speech band to a single sideband of a high-frequency carrier wave and then rectifying the resulting signal. An essential difference between these two methods is that in the first case both sum and difference

SPECTROGRAMS OF THE PHRASE "THERE ARE MANY CALL(S)"



OVERALL WEIGHTING +3db/OCTAVE.
MARKING RANGE 40 db
DURATION 0·9 SECS. (APPROX).

Fig. 2.

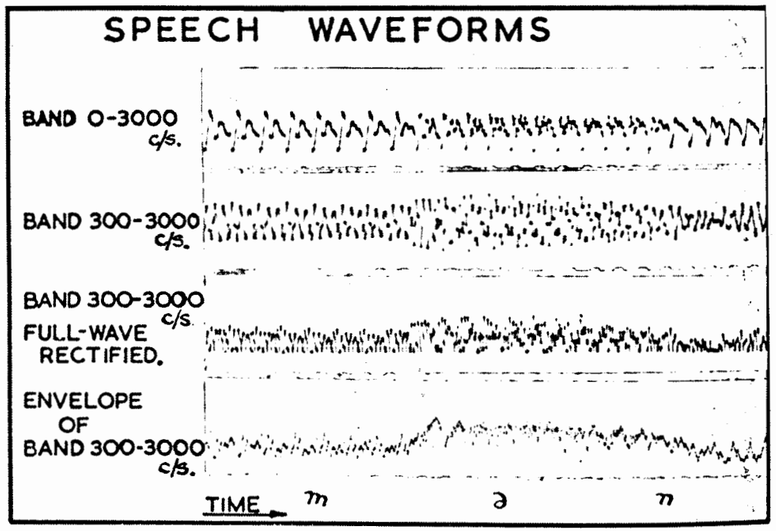


Fig. 3

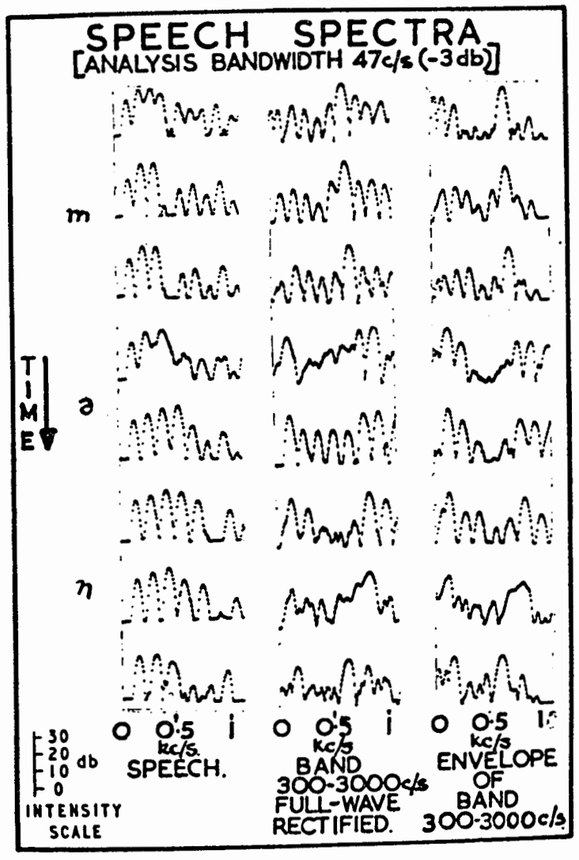


Fig. 4.

products fall within the original band whereas in the second case only the difference products contribute to the result. Each one of these methods can provide an accurate estimate of the fundamental glottal frequency of sustained vowels but it will be shown that difficulties arise when continuous speech is used.

The two lower traces of Fig. 3 show the waveforms that result from intermodulation of the band-limited waveform of /mən/ and Fig. 4 shows a sequence of spectral cross-sections of the signals that are produced by the two intermodulation processes and also corresponding spectral cross-sections of the original speech. It can be seen, from Fig. 4, that the fundamental can be deduced from the pattern of the peaks of the spectral cross-sections of the speech, even when the band of frequencies below 300 c/s

has been removed. The two intermodulated signals, however, appear to provide a rather less reliable indication of the glottal fundamental. When the vocal tract is changing rapidly, as for example during the first, third and eighth scans of Fig. 4, considerable signal energy appears in the region below the frequency of the glottal fundamental. At times the required glottal indication cannot be separated completely from this unwanted low-frequency signal. It may be possible to derive an approximate estimate by using an intermodulation scheme followed by a tracking filter which attempts to follow and select the fundamental. The present evidence, however, suggests that a more accurate estimate may be obtained by measuring the short-term spectral cross-section of the speech, determining the locations of the peaks of the spectrum and using the resulting simplified spectral pattern to estimate the fundamental glottal frequency.

Joint Speech Research Unit
Ruislip, Middlesex