Proceedings of the Institute of Acoustics

PITCH ESTIMATION FOR TWO OVERLAPPING VOICES

Zhao J. and Denbigh P.N.

University of Sussex, School of Engineering and Applied Science
Falmer, Brighton, BN1 9QT

1. INTRODUCTION

Many techniques have been described in the literature for extracting the time varying pitch of a single voice and an excellent overview is given by Hess [1]. The estimate in such a situation is usually accurate . When two or more voices are superposed however, the accuracy of the pitch estimate for each speech degrades rapidly [2,3]. This paper describes a reasonably successful attempt at determining the pitch of each of two overlapping speech signals which have different pitch ranges. The ultimate objective of the work is to produce a hearing aid that can enhance one speech relative to the other, using a comb filter to pass the harmonics of the target speech during voiced sounds while simultaneously rejecting much of the energy from the interfering speech. The emphasis of this paper however is on extracting the pitch of overlapping sounds, rather than on progress regarding this latter task of speech separation.

2. SOME BACKGROUND

One of the many methods of determining the pitch of a single speech signal is the recent one of subharmonic summation described by Hermer [4] and, since it forms an important part of the new method, a brief description of it is appropriate. The essential principle is to add to the speech spectrum a large number of frequency compressed speech spectra. The original spectrum is added with the speech spectrum compressed by two in frequency, the speech spectrum compressed by three in frequency, the speech spectrum compressed by four in frequency, and so on. If a periodic signal rich in harmonics is present, each of these modified spectra will contain a peak that lines up with the fundamental frequency of the periodic signal in the unmodified spectrum. For example the fundamental frequency in the unmodified spectrum will line up with the second harmonic in a spectrum that is compressed by two, and with the third harmonic in a spectrum that is compressed by three, and with the fourth harmonic in a spectrum that is compressed by four, etc etc. The line up of harmonics in this way causes a large peak to arise in the summed spectrum. For voiced speech signals this large peak occurs at the pitch frequency. Further details of the method are given by Hermes [4] where an important difference of detail is that a logarithmic frequency axis is preferred because the modified spectra are then obtained by simple shifts of log2, log3, log4 etc, rather than by compressions of 2, 3, 4 etc; also it is more accurate and computationally efficient. Using a logarithmic frequency axis the summed spectrum is termed the "subharmonic sum spectrum" (SHSS). The technique has been successful with single voiced sounds but cannot directly apply to the case of overlapping sounds. In the case of two overlapping speech signals for example, two peaks in the subharmonic sum spectrum need to be found, each related to the pitch of the corresponding voice. While the largest peak in the subharmonic sum spectrum may correspond to one pitch frequency, the next largest peak is unlikely to correspond with the second pitch frequency. This is primarily because harmonics of one voice can be harmonically related to some of the harmonics of the second voice, and some harmonics of the weaker voice can be masked by the harmonics of the other voice, with the consequent introduction of spurious peaks in the summed spectrum that are higher than the peaks for the weaker voice.

3. OUTLINE OF METHOD

The new method is summarised by the flow chart of Fig.1. It commences with short term spectra based on 40ms segments (or frames). It then seeks any rapid change in the spectrum that appears to correspond with the sudden onset of a sound. When this occurs the changing part of the spectrum can be separated from the static part. Both of these separated spectra are now individually examined by the subharmonic summation method. In each case the subharmonic sum spectrum is examined to determine whether it corresponds to a voiced or unvoiced sound. If voiced the new measurement of pitch is adopted as the best up-dated value. An unvoiced result, in contrast, is ignored.

The detection and exploitation of the onsets of sounds is an important ingredient of the new method. However the onset condition is likely to be very short in duration (by definition) and at other times an alternative strategy is needed. This alternative stategy is centred around the tracking of the pitch, i.e. on exploiting measurements of the sounds in previous frames.

At times not corresponding to onsets the second strategy begins by forming the subharmonic sum spectrum from the raw spectrum. The pitch of each sound is then estimated from this subharmonic sum spectrum in a way that depends on the analysis of this sound in the previous frame. If one sound has been deemed to be voiced in the previous segment, the pitch of this voice for the presentsegment is determined by considering only that feature in the subharmonic sum spectrum that is close to the previous pitch measurement of this voice. In effect the idea is to be able to follow movements of a relevant peak in the subharmonic sum spectrum, and thus to track pitch, even if that peak is weak in the subharmonic sum spectrum.

If the voice in the previous segment has been deemed to be unvoiced the highest peak in the summed spectrum is considered.

Two strategies have been described, one using onset information and one not. The main advantage of including the onset method is that it can enable the algorithm to "lock on" to a weak speech signal at its moment of onset in a way that would not be possible without it. It should be noted that the subharmonic summation method was originally intended to use only the strongest peak in the subharmonic sum spectrum - a concept that totally breaks down when two speech signals are present.

4. SOME DETAILS OF METHOD

4.1 Spectral Estimation

The signal is divided into overlapping segments (or frames), 20ms apart and 40ms long. Each segment is weighted with a Hamming window , appended with 11.2 ms of zeros and the amplitude spectrum is evaluated using the FFT. This spectrum will typically contain large and weak peaks caused by the harmonics of voiced sounds, plus a background due to unvoiced sounds and other effects. The most significant peaks are extracted by rejecting those parts of the spectrum that lie below either of two amplitude thresholds. One of the thresholds follows the general slow spectral density variation across the frequency band of an average voiced excitation. It is chosen to be very low in amplitude and the objective of this threshold test is to eliminate any features that are too weak to be of reliable significance. The second amplitude threshold is one that varies more rapidly across the frequency band. It is derived from the measured spectrum itself, using a version of it that is only slightly smoothed. This second threshold is therefore high in the neighbourhood of large peaks. Its purpose is to eliminate peaks that are unreliable because of their close proximity to larger peaks, caused for example by sidelobe effects.

4.2 Onset Detection And Spectrum Separation At Onset

The onset is detected by measuring the spectrum change. Spectrum changes are measured by a process involving several frames of the time varying spectrum but similar in principle to subtracting neighbouring spectra. If no other voice is present at the moment of onset the technique is unnecessary. If one voiced sound is already present however the sudden increase of the "differential" spectrum will indicate the onset of a second voiced sound. By comparing the peaks of the differential spectrum with those of the raw spectrum the various peaks can be attributed to one voice or the other, i.e. separated.

4.3 Validity Check of a Peak In Subharmonic Sum Spectrum

It is always possible to identify the highest peak in the subharmonic sum spectrum (or in the part of subharmonic sum spectrum close to the previous measurement if tracking is used). This highest peak, however, does not always indicate the pitch of a voice. Therefore the association of this peak with the fundamental frequency of a voiced sound is checked by demanding that

a) its amplitude should be a reasonable fraction (typically 0.3) of the sum of all the peaks in the raw spectrum and

b) the corresponding pitch estimate falls within the pre-defined pitch range of this voice.

Test (a) ensures that the selected peak arises from a voiced sound, and test (b) ensures that any unreliable pitch estimate for this voice is removed.

If the peak in the subharmonic sum spectrum sought for one sound is not valid, this sound is classified as unvoiced in this segment. Thus a voiced/unvoiced decision is made through such validity check.

5. EXPERIMENTAL RESULTS

Speech signals were recorded directly to the hard disc of a PC containing a plug-in Metrobyte DAS-20 A/D and D/A converter board. The system used a sampling frequency of 10kHz and direct memory access. One of the speech signals recorded was the sentence "The famous event that took place here was the murder of Archbishop Thomas Becket in 1170 by Henry II's knights" spoken by a male speaker. The section of speech underlined is shown in Fig.2a, and Fig.2c shows the pitch variation of this section as determined by the conventional subharmonic summation method. A second speech signal was the sentence "My problem is with my mother, who is now well over seventy, a widow." spoken by a female speaker The section of speech underlined is shown inFig.2b, and Fig.2d shows the pitch variation of this section as also determined by the subharmonic summation method. An independent confirmation of the accuracy of the pitch measurements has been obtained by using one of the measurements to control the fundamental frequency of a comb filter, through which the combined speech signals are then passed. The output was a low distortion speech signal.

The next step was to superpose the two conversations on the computer and then to apply to this the procedure outlined in this paper to attempt to extract the pitch of both conversations separately. Fig.2e shows the result for the male voice, and Fig.2f that for the female voice. It is seen that these are generally in good agreement with the pitch variations shown in Figs.1c and 1d.

6. CONCLUSIONS

Attempts to extract the pitch of two overlapping conversations with similar intensity have been reasonably successful so long as the pitch ranges of two voice are different . A major ingredient of the method is the recognition of the onset of a voiced sound on the basis of a spectral change. The relevant spectral features are then appropriated to that sound so that its pitch may be determined, even though another voiced sound may be present at the same time. Thereafter the pitch is tracked until voicing ceases. The tracking algorithm improves the ability to estimate the pitch of the weaker voice and reduces the likelihood that interaction effects between the two voiced sounds should cause a false determination of pitch.

Work is proceeding on using the pitch measurements to control comb filters that will enhance a target conversation when other unwanted conversations are simultaneously present.

7. REFERENCES

[1] W Hess, 'Pitch Determination of Speech Signals', Springer-Verlag (1983)

[2] T W Parsons, 'Separation of Speech From Interfering Speech by Means of Harmonic Selection', J. Acoust. Soc. Am. 60, 911-918. (1976)

[3] R J Stubbs & Q Summerfield, 'Algorithms for Separating the Speech of Interfering Talkers; Evaluation With Voiced Sentences, and Normal-hearing and Hearing-impaired listeners'. J. Acoust. Soc. Am. 87, 359-372. (1990)

[4] D J Hermes, 'Measurement of Pitch By Subharmonic Summation', J. Acoust. Soc. Am. 83, 257-264 (1988)

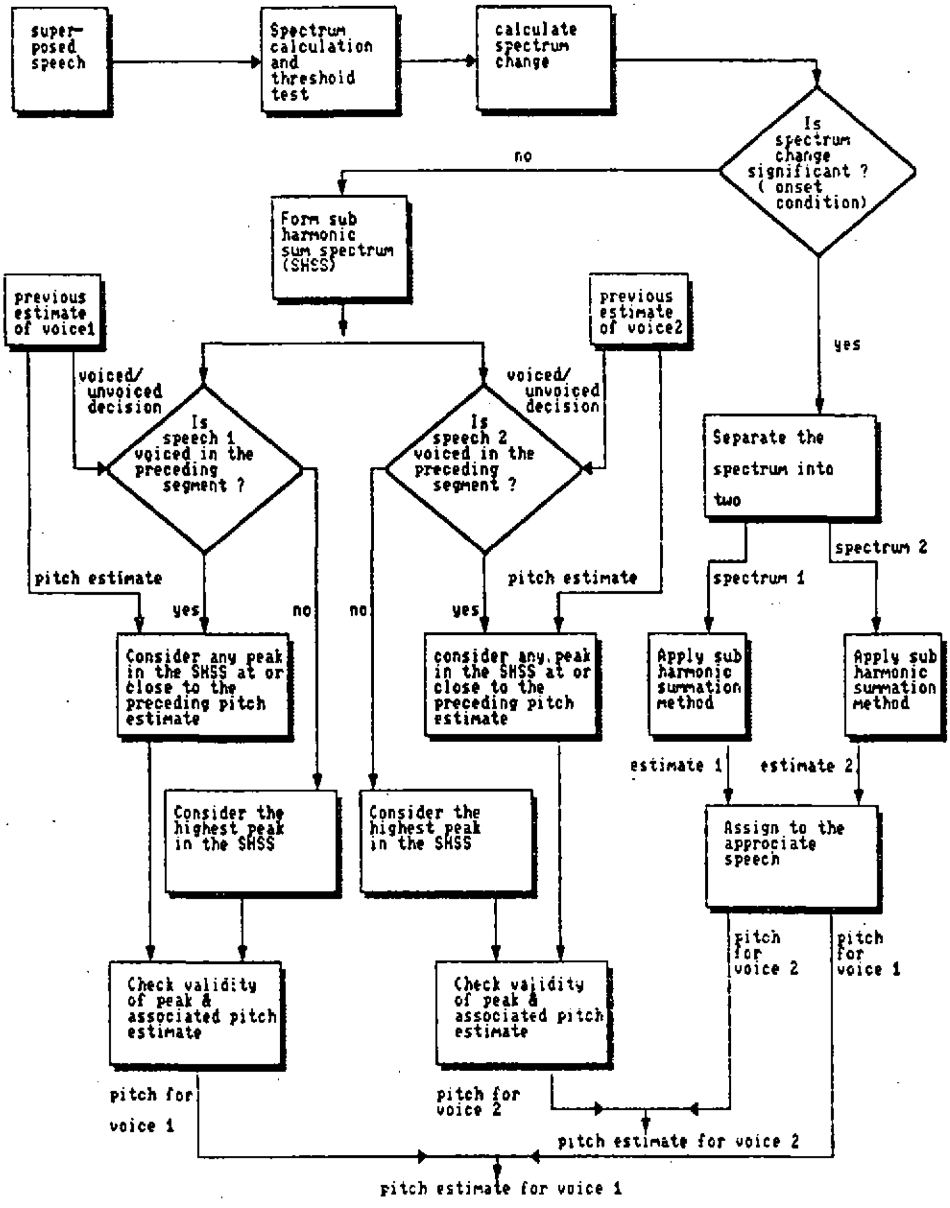

FIG.1 flow chart of algrithm

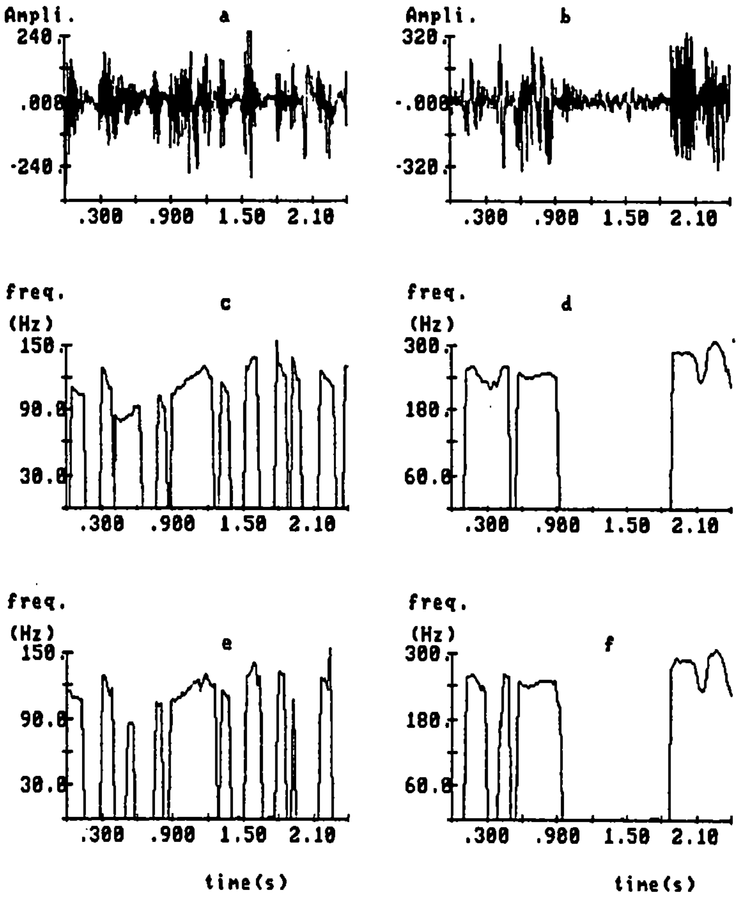

FIG.2 Time waveform and pitch estimates of speech signals, a) an c) male voice alone, b) and d) female voice alone. e) and f) pitch estimates when the speech signals are superposed.