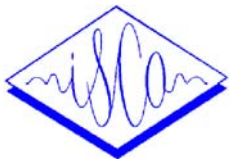


5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

ROBUST MEASUREMENT OF FUNDAMENTAL FREQUENCY AND DEGREE OF VOICING

John N. Holmes

Speech Technology Consultant, 19, Maylands Drive, Uxbridge, UB8 1BH, UK
Tel. +44 1895 236328, E-mail: jnh@jnholmes.demon.co.uk

ABSTRACT

Both for robust fundamental frequency (F0) measurement and to provide a degree of voicing indication, a new algorithm has been developed based on multi-channel autocorrelation analysis. The speech is filtered into eight separate frequency bands, representing the lowest 500 Hz and seven overlapping band-pass channels each about 1000 Hz wide. The outputs of all the band-pass channels are full-wave rectified and band-pass filtered between 50 Hz and 500 Hz. Autocorrelation functions are calculated for the signals from all eight channels, and these functions are used both for the F0 measurement and for the voicing indication. Optional dynamic programming is provided to maximize the continuity of position of the correlation peaks selected for fundamental period measurement. The algorithm has been designed for coding onto a 16-bit integer DSP, using less than 4 MIPS processing power and 1500 words of data memory.

1. INTRODUCTION

There are many potential applications for speech excitation analysis, such as low-bit-rate speech coding, talker verification, language identification and use of prosody in automatic speech recognition. For most of the time the fundamental frequency (F0) can be represented by a slowly varying parameter, although in speech production this simple model of vocal fold action is often seriously inaccurate. Vocal fold vibrations are sometimes irregular; sometimes voicing periods are alternately longer and shorter; sometimes the frequency suddenly changes by a factor of 2:1, thus making it appear that every alternate period has been omitted. It is therefore useful if the F0 parameter can also be made to represent irregularities in period when they occur. In addition it is desirable that the F0 analysis should be as robust as possible in the presence of added noise, reverberation and various other distortions.

For most speech sounds the excitation is clearly either all voiced or all voiceless. However, in breathy voice or voiced fricatives the excitation appears periodic in the lower part of the audio spectrum, but has a substantially random structure at higher frequencies. Many workers are therefore now showing interest in providing a parametric representation of excitation mixture as well as F0, for puproses such as automatic speech recognition or to improve the quality in low-bit-rate speech coding [1,2].

This paper is concerned with providing a very robust F0 parameter and a variable degree of voicing to use for any of the applications and types of speech signal described above, and in particular to drive the formant synthesizer described in [3]. The two output parameters are determined at 100 frames/s.

2. PRINCIPLES OF NEW ALGORITHM

2.1. F0 estimation

Spectrograms. When a wideband spectrogram of speech is examined visually, it is usually obvious to an experimental phonetician where excitation from glottal closure has occurred. During breathy voice or voiced fricatives, the glottal closures are still apparent in the lower-frequency parts of the spectrum, although the structure appears random in the higher-frequency regions. High levels of added noise do, of course, mask the structure of the lower energy regions of the spectrum, but in any regions where the formants are intense enough to show above the noise, the excitation structure will still be apparent. It is thus possible to estimate the fundamental frequency even when a speech signal is so corrupted by noise that on the spectrogram only one formant is visible above the noise level.

Autocorrelation. To estimate F0 one needs to measure the spacing of successive excitation pulses. Short-time autocorrelation is a well-established method for this purpose and, provided the time window is sensibly chosen, it can be used to analyse periodic signals of slowly varying fundamental frequency. However, it is important to choose the most suitable signals for analysis. The speech waveform by itself is not suitable, because periodicity at formant frequencies can often cause more prominent correlation peaks than F0 periodicity, particularly when F0 is changing. From the discussion in the previous paragraph, it can be seen that a suitable candidate might be the signal envelope within bands wide enough to carry the envelope fluctuations at the highest required value of F0. Because of noise and other possible distortions there is no guarantee that the envelope of all such bands will contain a periodic signal. It is necessary, therefore, to calculate several autocorrelation functions of separate bands, such that it is guaranteed that any formants which show periodicity will be included in at least one of them. The separate autocorrelation functions must then combined with most weight given to those showing periodicity most clearly.

The idea of using the envelopes in multiple bands is not new. Nearly 40 years ago Gill [4] developed a method using a combination of four channels of autocorrelation derived from different parts of the spectrum. His method worked very well considering the limitations of electronic hardware at that time. More recently Walliker and Howard [5] used the envelopes of nine spectral channels as inputs to a multi-layer perceptron voice pulse detector. In the current paper the novelty of the algorithm is in the details of the correlation measurement and the method of combining multiple channels to derive estimates of both the fundamental frequency and degree of voicing.

Correlation calculation. The speech is filtered into eight frequency bands, representing the lowest 500 Hz and seven band-pass channels, with centre frequencies spaced at 500 Hz intervals from 500 Hz to 3500 Hz. The band-pass channels have been made 1000 Hz wide so that they can contain sidebands of amplitude modulation for F0 up to 500 Hz. The outputs of all the band-pass channels are full-wave rectified and then band-pass filtered between 50 Hz and 500 Hz, so that the signals then represent the envelope ripple within each channel.

Autocorrelation functions are calculated for the signals from all eight channels. The input is multiplied by a bell-shaped window whose half-amplitude width is 30 ms, and a maximum lag of 20 ms was chosen to allow F0 values down to 50 Hz. After normalization the correlation functions are multiplied by a lag-dependent weighting factor that takes into account the window shape and duration, so that for a perfectly periodic signal the height of the peak at the fundamental period is independent of the period.

Channel combination. The eight separate autocorrelation functions are combined with appropriate weights before determining the peak to represent signal periodicity. Although formant periodicity in the baseband channel can often produce higher peaks than the fundamental periodicity, the fundamental-period peak is usually sharper than those from the envelope channels, so allowing a more precise F0 measurement. It is therefore still almost always advantageous to give this channel a significant weight in the correlation mixture.

Any envelope autocorrelation functions which are useful in determining the periodicity will be fairly similar in shape, so the cross-correlation coefficients between all pairs of these functions are calculated. The channel which has the highest sum of cross-correlation coefficients with all other channels is identified as the most important channel for the weighted sum, and all other channels which correlate well with this channel are also included. Sometimes, however, particularly in noisy speech, none of the envelope channels will correlate well with any other, but even then it is possible that one channel may have a sufficiently periodic signal to be useful. In this case the height of the highest peak of the autocorrelation function of that channel can be used to determine a channel-weighting factor.

Selection of fundamental period. In the combined autocorrelation function the highest peak is usually at the fundamental period, but occasionally transient effects cause the period peak to be lower than earlier or later peaks. When several consecutive peaks are of about the same height, the earliest peak is the required one, so a rule has been included to ensure this peak is chosen in these cases. Occasionally, at the onset of voicing for high-pitched female speakers, the large level-transient causes a low-frequency input to the correlation calculation. The first correlation peak, at the fundamental period, is then sometimes much lower than subsequent peaks. However, after two frames this transient will have decayed sufficiently for the correct peak to be chosen. It is not possible to know, for the first frame of such a voiced sound, whether the cause of the lower first peak is the level transient or a true longer period. Dynamic programming with two frames delay is therefore provided as an option, to maximise the continuity of chosen peak positions from frame to frame.

2.2. Degree of voicing

During normal speech production, whenever mixed excitation occurs the lower frequency region is always voiced, and the random excitation is not usually apparent below 2 kHz or even higher. This situation can be well modelled by choosing a degree-of-voicing indication which controls the frequency above which the voicing mixture starts to change, as advocated in [3]. In the current algorithm, for any of the envelope channels which indicate a clear periodicity, the excitation is classified as voiced at the corresponding frequency. By implication it will also be voiced at all lower frequencies, even if the periodicity is hidden by poor signal-to-noise ratio.

If the correlation in higher channels does not show periodicity, there are two possible explanations. One is that there is genuine turbulent excitation in these channels, and the other is that aperiodic background noise exceeds the speech signal level. If the level in higher channels is comparable with the level of any channels for which voicing is detected, no algorithm could distinguish between these two situations, and it seems best to classify these frames as carrying mixed excitation. However, if the signal level in the high frequency channels is low, the periodicity may not be detected simply because the speech signal level is below a very low background noise level. This latter situation occurs frequently in high-quality speech when voicing is decaying into silence, particularly for sounds with mainly low-frequency energy. In these circumstances it is more realistic to classify the signal as fully voiced.

The algorithm deals with mixed excitation by determining the highest-frequency channel for which there is clear periodicity, and classifying that channel and all lower channels as voiced. It then compares the signal level with that in successive higher channels, which are marked as voiced until a channel is found that does not give a cumulative level reduction of at least 4 dB per channel. After each channel has been given a voiced or voiceless marking in this way, the final degree of voicing indication is derived by counting the number of voiced markings, so giving a scale between 0 and 8. This scale is finer than is necessary, and even during steady conditions of mixed excitation it is not unusual for frame-to-frame variations of one or two levels to occur. A more realistic voicing indication is therefore derived by median smoothing of the 9-level parameter.

3. IMPLEMENTATION

Because of the likely use of this algorithm in real-time speech processing systems, care has been taken to make it capable of being coded efficiently on a low-cost integer DSP chip. The core of the algorithm is all coded in ANSI standard C, using integer arithmetic throughout with 16-bit variable storage. A precision of 32 bits is used for multiply-accumulate functions, and for the dividend of the very few division operations. The input signal is linearly coded with 16 bits at a sampling rate of 8 kHz, and a block-scaling technique is used over the duration of the correlation windows to ensure that there is no significant loss of accuracy for very low-level signals.

After the channel signals have been limited to 500 Hz, they are down-sampled to 1600 Hz to reduce the computation and

memory requirements of the correlation calculations and the remainder of the algorithm. However, the consequent coarse quantizing of the correlation lag makes it necessary to do parabolic interpolation through three points round each correlation peak to derive both peak height and position with increased accuracy. Although this interpolation gives a fairly small error in peak position, for very short periods the error can still give a significant proportional error in F0. To reduce this error the correlation function is scanned to the end to look for significant peaks whose positions are close to multiples of the selected period. The position of the last such peak is then divided by the position ratio to determine the period with much higher accuracy. The interpolation error for high-pitched speech is typically reduced by a factor of between 5 and 10 by this means.

It is estimated that the complete algorithm, if coded for real-time use onto a 16-bit integer DSP, would require less than 4 MIPS processing power and 1500 words of data memory.

4. PERFORMANCE

The algorithm has been tested on a wide variety of speech signals, including some with non-linear distortion and very high level background noise, and several with problems caused by speech production (e.g. wide pitch range, rapid pitch changes, frequent low-pitched irregular voicing, breathy voice, etc.).

In common with other excitation analysis algorithms, for most of the time the output is obviously correct. Problems arise mainly when the excitation does not fit the assumed parameter-ization model. In these cases there can be no “correct” result, and one can only judge whether the output gives a plausible interpretation within the limitations of the model. The performance has therefore been assessed by visually examining all potential places where the results might be in error, and looking to see if the derived excitation parameters were consistent with the signal properties as interpreted by a human expert looking at a wideband spectrogram. A program has been written that displays all places where either F0 or degree of voicing changes significantly from frame to frame, and automatically aligns the speech spectrogram with the display. Inconsistencies are easily seen by this means.

Only very rarely have the results disagreed with the best interpretation that could be given by the human expert. Sensible F0 measurement has been possible even for sounds so contaminated by noise that the only unambiguous spectro-graphic evidence of voicing periodicity was above 2 kHz.

Fig. 1 shows 500 Hz or 1000 Hz spectrograms of speech signals which present special problems for excitation analysis, aligned with F0 and voicing parameters. To the right of each spectrogram is the set of 8 correlation functions for one marked frame of data, with the weighted combination displayed above. To make it easy to judge the measurement accuracy, the F0 parameter has been used to synthesize an excitation pulse waveform, with the pulse positions marked as dotted lines on the spectrograms. As the position of the first pulse of each pulse train is arbitrary, it has been manually adjusted to make the best average alignment with the spectrogram features.

Fig. 1a shows a voiced fricative, [z], with a total of 13 frames of mixed excitation before the vowel. The correlation functions, for one of the frames with weaker voicing, only show clear periodicity in the lowest two channels, but low level at 1000 Hz has also caused the third channel to be marked as voiced.

Fig. 1b shows an interval between the words “Hello operator”, during which the voice went creaky for 4 frames. Although the correlation analysis was not able to follow all the irregular pulse timing, the regenerated excitation pulses are at approximately the right spacing during the creak. This is clearly seen in the correlation functions for the marked frame. Maximum voicing was not registered because the envelope shapes in the top three channels were different from those of the lower channels.

Fig. 1c illustrates that the algorithm works at very high F0 (465 Hz), and shows the effect of gradual decay of frication as the [s] merges into the following vowel. The correlation functions show that transients in the first frame of voicing disturb the correlation function shape for the 500 Hz channel.

For most of Fig. 1d alternate glottal closures are only very weakly exciting the vocal tract. The F0 values given, although only half the frequency of the true vocal fold vibration, make a reasonable attempt at synthesis of the very-low-pitched sound of the natural speech.

In Fig. 1e the noise and distortion are almost completely masking the periodicity for the last few voiced frames, but it can be seen that the correlation structure between 2000 and 2500 Hz is still sufficient for the correct F0 to be measured.

5. ACKNOWLEDGEMENT

Thanks are due to the Speech Research Unit of DERA (UK) for providing the speech files for Figs. 1d and 1e.

6. REFERENCES

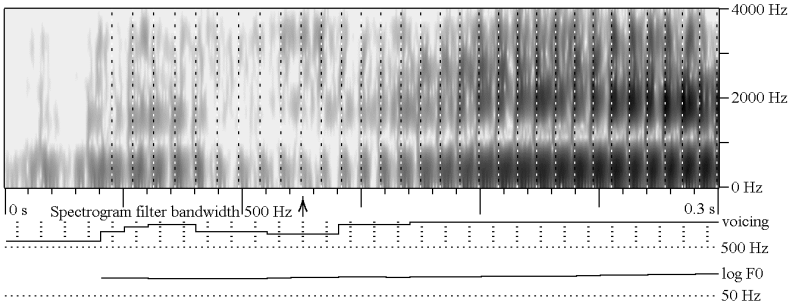
1. McCree, A., and De Martin, J.C. "A 1.7 kb/s MELP coder with improved analysis and quantization" *Proc IEEE ICASSP*, Seattle, 123-126, 1998.

2. Cho, Y.D., Kim, M.Y., and Kim, S.R. "A spectrally mixed excitation (SMX) vocoder with robust parameter determination" *Proc IEEE ICASSP*, Seattle, 127-130, 1998.

3. Holmes, J.N. "A parallel-formant synthesizer for machine voice output" In Fallside, F. and Woods, W.A. (Eds) *Computer Speech Processing*, Prentice Hall, 163-187, 1985.

4. Gill, J.S. "Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods", In Cremer, L. (Ed.) *Proceedings of the Third International Conference on Acoustics Stuttgart 1959*, Elsevier, 217-224, 1961.

5. Walliker, J. R., and Howard, I. "Real-time portable multi-layer perceptron voice fundamental-period extractor for hearing aids and cochlear implants" *Speech Communication*, Vol. 9, 63-72, 1990.



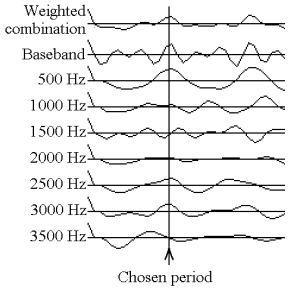
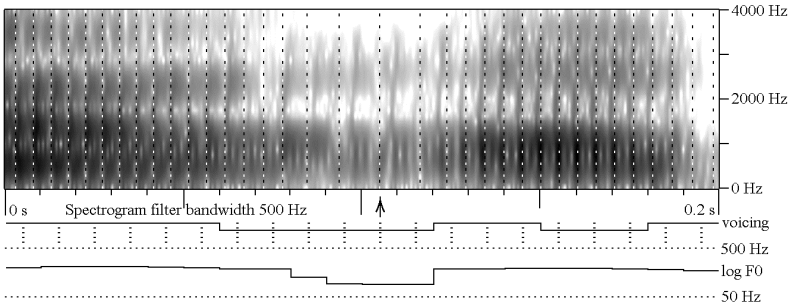


Fig. 1a.

Telephone speech, start of the word "zero", spoken by a male speaker.



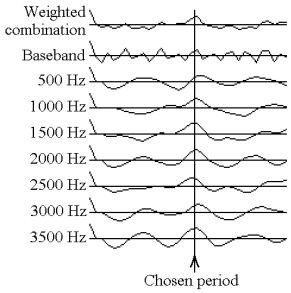
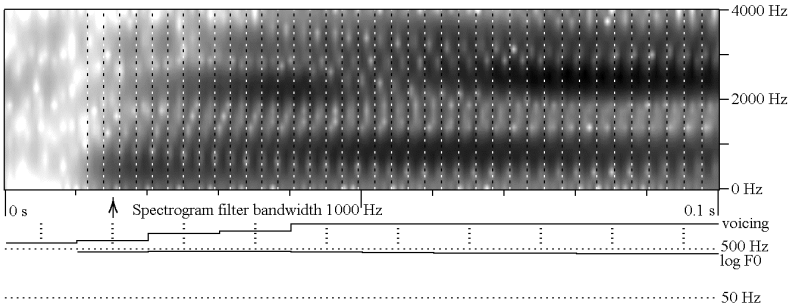


Fig. 1b.

"Creaky voice" between two words of high quality male speech.



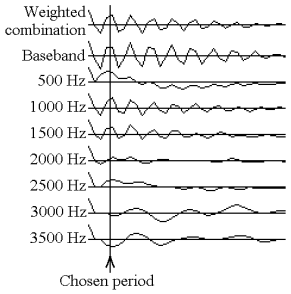
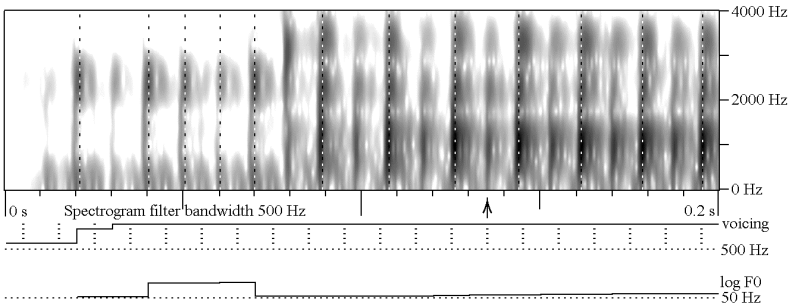


Fig. 1c.

Start of voicing of the word "say" spoken by a high-pitched female speaker.



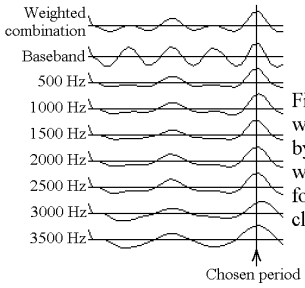
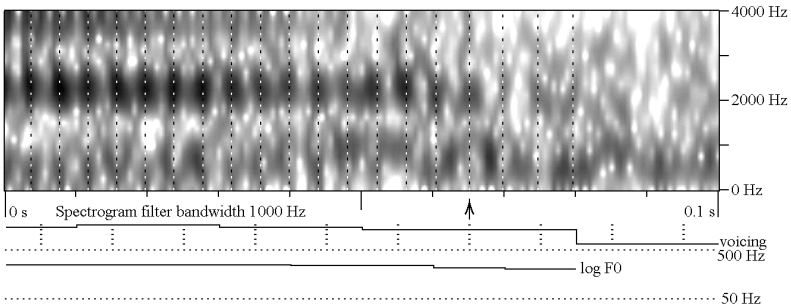


Fig. 1d.

First half of the word "nine" spoken by a male speaker with different mode for alternate glottal closures.



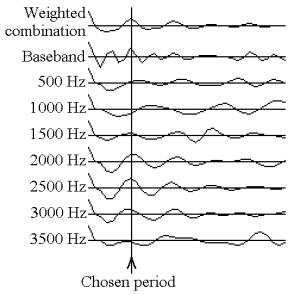


Fig. 1e.

Last part of the word "three" recorded in a very noisy military vehicle by a male speaker.

Fig. 1. Spectrograms of various speech fragments, with correlation analyses of frames marked by arrows. See text for details.