PERFORMANCE AND OPTIMIZATION OF THE SEEVOC ALGORITHM

Weihua Zhang and W. Harvey Holmes

School of Electrical Engineering, The University of New South Wales, Sydney 2052, Australia
wzhang@iowave.com, h.holmes@unsw.edu.au
http://www.ee.unsw.edu.au

ABSTRACT

In most low bit rate coders, the quality of the synthetic speech depends greatly on the performance of the spectral coding stage, in which the spectral envelope is estimated and encoded. The Spectral Envelope Estimation Vocoder (SEEVOC) is a successful spectral envelope estimation method that plays an important role in low bit rate speech coding based on the sinusoidal model.

This paper investigates the properties and limitations of the SEEVOC algorithm, and shows that it can be generalized and optimized by changing the search range parameters α and β. Rules for the optimum choice of α and β are derived, based on both analysis and experimental results.

The effects of noise on the SEEVOC algorithm are also investigated. Experimental results show that the SEEVOC algorithm performs better for voiced speech in the presence of noise than linear prediction (LP) analysis.

1. INTRODUCTION

Spectral envelope estimation is one of the most important aspects of speech processing, especially in speech coding, where the quality of the decoded synthetic speech is very dependent on the performance of the spectral estimation method.

For low bit rate coders, the sinusoidal representation of the speech spectrum has been proposed as one of the most promising approaches [1, 2, 3]. One of the main methods used for estimating the spectral envelope in sinusoidal coders is the Spectral Envelope Estimation Vocoder (SEEVOC) algorithm [1], which has some desirable properties compared with the alternatives. A very useful by-product of the algorithm is an excellent pitch estimation algorithm.

The SEEVOC algorithm is basically a nonlinear method of estimating the spectral envelope of the speech signal in the frequency domain. It works by interpolating between major peaks of the spectrum which are separated at approximately the correct pitch interval. It is therefore not influenced by low-level peaks, which may be due only to noise or processing artefacts such as sidelobe leakage, but which would influence simpler peak interpolation schemes.

In spite of its importance for coding, there does not appear to be any analysis of the performance of the SEEVOC algorithm in the available literature, nor any attempt to optimize it. In this paper we analyze the behaviour of this algorithm, both theoretically and experimentally. Theories are developed that allow us to effectively optimize the algorithm and reduce the effects of inaccurate preliminary pitch estimates. Further, its robustness in noise is analyzed and compared with the use of LP analysis.

2. THE SEEVOC ALGORITHM

The first step of the SEEVOC algorithm is to calculate the magnitude of the short-time Fourier transform (STFT) of the speech frame.

The SEEVOC algorithm also requires as input an initial estimate *CP* of the pitch, called coarse pitch. Since it is generally believed that the exact choice of *CP* is not critical, the emphasis is usually on simplicity for the determination of *CP* – e.g. a simple pitch determination algorithm (PDA) could be used, or the pitch in the current frame may be assumed to be the same as that found in the previous frame.

The algorithm works by searching in the frequency domain for major spectral peaks ω_k and their associated amplitudes A_k, $k = 1, 2, \ldots$, as follows. The search ranges for each major spectral peak depend on the lower frequency peaks already found and on *CP*. The first peak is found by searching the frequency interval $[\alpha, \beta]$ for the largest spectral value. In the original algorithm the constants are chosen to be $\alpha = CP/2$, $\beta = 3CP/2$. Suppose the amplitude and frequency of the first peak are (A_1, ω_1). Subsequent peaks (A_k, ω_k), $k = 2, 3, \ldots$, are then found by successively searching the intervals $[\omega_{k-1} + \alpha,\ \omega_{k-1} + \beta]$ for their largest spectral values, until the edge of the speech bandwidth is reached. If no true peak is found in a search region, then the amplitude of the largest end point is used and placed at the bin centre $\omega_{k-1} + CP$, from which the search procedure is continued. This search procedure is illustrated in Fig. 1.

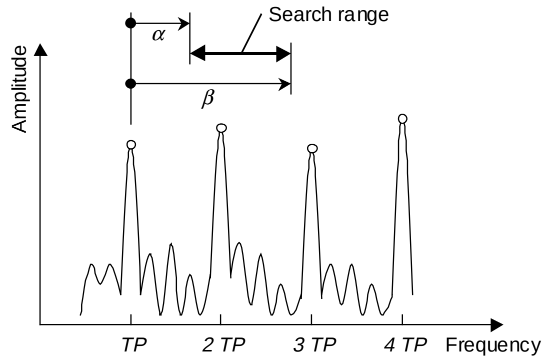

Fig. 1. The SEEVOC algorithm, illustrating the search for the second peak. *TP* denotes the true pitch, *CP* the coarse pitch – with luck, *CP* = *TP*. In the original algorithm the search range parameters are chosen to be $\alpha = CP/2$, $\beta = 3CP/2$.

Finally, the SEEVOC envelope is obtained by connecting all the chosen peaks by linear or cubic spline interpolation in the log-amplitude domain. This is the SEEVOC estimate of the system amplitude (or envelope) response $A(\omega)$. The result is illustrated in Fig. 2.

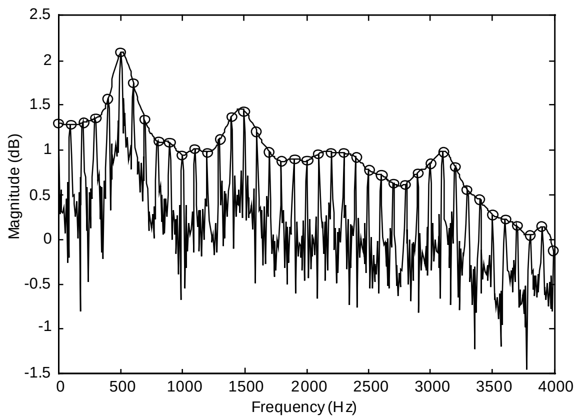

Fig. 2. Speech spectrum with SEEVOC peaks (circles) and SEEVOC envelope (using cubic spline interpolation).

The SEEVOC search strategy forces the selected peaks to be separated at approximately the correct pitch interval. The expectation is that these peaks will be approximately harmonics of the fundamental frequency if the speech is voiced. A major advantage of this peak selection method is that it ignores all low-level peaks, which may be due only to noise or processing artefacts such as sidelobe effects, in favour of the largest peaks in each harmonic interval. It is also relatively tolerant of inaccurate pitch estimates *CP*.

3. THE EFFECT OF COARSE PITCH

In order to examine how well the SEEVOC algorithm estimates the envelope of a speech signal, it is necessary to use test signals with known spectral envelopes and pitches. Accordingly, synthetic voiced speech signals were used, generated by applying trains of impulses to vocal tract filters derived from actual speech signals by LP analysis. Thus the frequency response of the vocal tract filter is the true envelope of the synthetic speech signal. This allows exact comparison of the SEEVOC envelope with this true envelope.

Spectral Distortion (or difference) between the true envelope and the SEEVOC envelope was used as the objective measure. This is calculated over a fine grid of frequency points by the formula

$$SD = \left\{ \frac{1}{L} \cdot \sum_{i=0}^{L-1} \left(10 \log \frac{PX_i}{K\ PY_i} \right)^2 \right\}^{\frac{1}{2}},$$

where L is the number of frequency points, PX_i is the power spectrum at the i-th frequency, and PY_i is the corresponding SEEVOC value. The gain factor K, which is included to allow for the fact that the scales of the two spectra may be quite different, is chosen to minimize *SD*.

If the *CP* = *TP*, the SEEVOC algorithm tends to select the true peaks of the signal, and the SEEVOC envelope should exhibit the minimum spectral distortion.

We ran experiments in which *TP* was held fixed while *CP* varied over a range. For each coarse pitch the spectral distortion was calculated. The result of a typical case is shown in Fig. 3. In this example the true pitch is 12.8 frequency samples, which corresponds to 100 Hz.

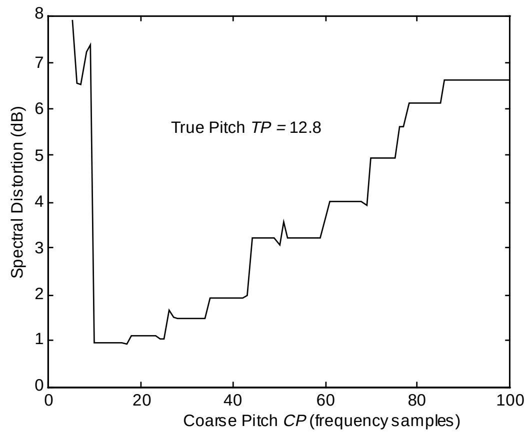

Fig. 3. The effect of coarse pitch on the SEEVOC envelope (search parameters $\alpha = CP/2$, $\beta = 3CP/2$). Each frequency sample corresponds to 4000/512 = 7.8125 Hz.

In this example the minimum spectral distortion is about 1 dB, which is typical, and this is obtained when the coarse pitch is between 10 and 17 samples. The spectral distortion increases steadily (with jumps, but gracefully) when the coarse pitch increases further. It is very noticeable that the spectral distortion rises sharply as the coarse pitch decreases from 10 to 9 – this occurs because the search range then just misses the second and many subsequent peaks, so that the SEEVOC envelope is quite different from the correct one.

Hence it is important that the coarse pitch is not substantially underestimated, whereas the SEEVOC algorithm is relatively tolerant of overestimated coarse pitches.

4. IMPROVEMENTS OF THE SEEVOC ALGORITHM

From the foregoing, it is very important to choose the coarse pitch accurately, otherwise large distortions will occur. Our aim is to reduce the sensitivity of the SEEVOC algorithm to the choice of coarse pitch by selection of α and β. In addition, if the search range can be narrowed by this selection, the computational effort will also be reduced.

We can theoretically determine some simple bounds on the coarse pitch range for minimum error from the following considerations.

To do this we have to consider the relation of *CP* to *TP*. It is useful to introduce the concept of a confidence interval for the true pitch *TP*. Thus, we consider the confidence interval ($m\,CP$, $M\,CP$) in which *TP* may reasonably be expected to lie (e.g. with 99.9% confidence), where $0 < m < 1 < M$. The actual values of m and M depend on the nature of the initial pitch estimator. Accurate pitch estimators will have both m and M close to 1, whereas pitch estimators prone to pitch halving (doubling) will have $M > 2$ ($m < 1/2$).

There is actually a hierarchy of goals for the peak search. The primary goal is that the search range (α, β) should contain at least the next harmonic peak (plus possibly others). A secondary goal is that the search range should contain *only* the next peak – i.e. it should not contain the second or later peaks. If the primary goal is not achieved, the error may be very large. If the secondary goal is not achieved, there will occasionally be errors because of wrong peak selection, but these errors would normally not be as large as if the primary goal is not achieved..

From Fig. 1, to achieve the primary goal we need $\beta > M\,CP$ and $\alpha < m\,CP$ (i.e. the search range should contain the confidence interval for the true pitch). To achieve the secondary goal, we need $\beta < 2m\,CP$. Both goals can be met if it is possible to choose α and β such that

$$M\,CP < \beta < 2m\,CP\,,$$
$$0 < \alpha < m\,CP\,.$$

Both goals can be satisfied for β only if $M < 2m$, which is not always possible (e.g. if pitch doubling or halving can occur). Otherwise, the secondary goal $\beta < 2m\,CP$ must be sacrificed and the optimum choice of β is $M\,CP$, which at least (just) satisfies the primary goal.

However, there is no such problem with the choice of α. The lower bound of α is arbitrary, based on both this theory and our experimental results. But a lower limit (say 0.7 $m\,CP$) should be set mainly to limit the amount of computation.

If the PDA is accurate (m and M close to 1), both α and β can be chosen close to *CP*, which can greatly reduce the computation required. However, to cater for non-voiced as well as voiced frames the usual choice ($\alpha = CP/2$, $\beta = 3\,CP/2$) is a good one.

On the other hand, if the PDA is prone to doubling or halving, more care is required in the choice of α and β, based on the above inequalities. In such cases, significantly improved overall performance can be obtained, as shown in Fig. 4.

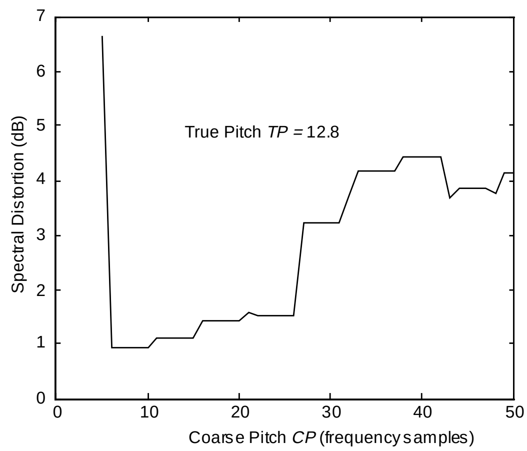

Fig. 4. The effect of coarse pitch on the SEEVOC envelope with α = 0.2 *CP*, β = 2.4 *CP*. This choice would be suitable if the PDA were prone to both pitch halving and pitch doubling problems.

5. THE SEEVOC ALGORITHM IN THE PRESENCE OF NOISE

It has been claimed [1] that the SEEVOC algorithm attains a degree of acoustic noise robustness by keying on the spectral peaks and ignoring the low level components, which are more affected by noise.

To examine this proposition, experiments were performed to investigate the effects of choice of coarse pitch on the SEEVOC algorithm in a noisy environment. The added noise was white Gaussian, and the input SNR was varied over the range 0~30 dB. The spectral distortion curves in a typical case are shown in Fig. 5.

For SNR above about 30 dB the performance is hardly affected, but at low SNRs the spectral distortion deteriorates, as expected. However, the choice of coarse pitch becomes less critical at low SNRs! That is, the optimum search range can be increased at low SNRs.

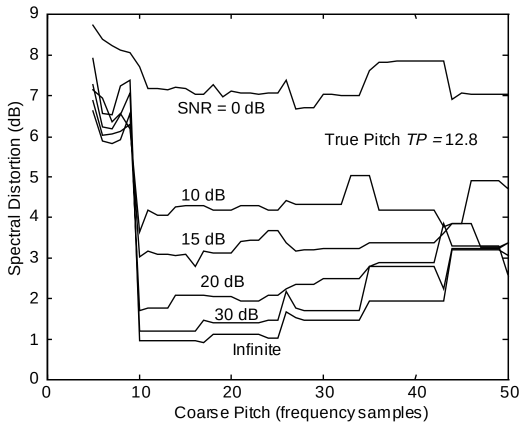

Fig. 5. The effect of white Gaussian noise on the SEEVOC algorithm.

Clearly, because the SEEVOC algorithm selects spectral peaks in all frequency ranges, the increase of spectral distortion at low SNRs comes from the low amplitude sections of the spectrum, which are most likely to have their peaks affected by the noise.

The performance in noise of the SEEVOC algorithm was also compared with that of linear prediction (LP) analysis, which is widely used as a method of spectral envelope estimation. In these experiments the autocorrelation method of LP analysis was used. The order p of the linear predictive filter can be used to control the degree of smoothness of the resulting spectrum. Since it is known that $p = 10 \sim 12$ is a good choice for speech [4], we chose $p = 10$. The obtainable spectral distortions with SEEVOC and LP was measured with true pitches in the range 50 to 400 Hz.

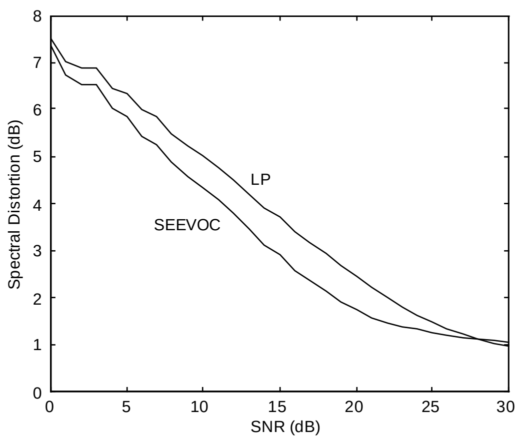

Fig. 6. Comparison of spectral distortions obtained from LP analysis and SEEVOC (true pitch $TP = 100$ Hz).

Figure 6 shows a typical result. For each true pitch period tested, the SEEVOC algorithm performs better than LP analysis in the approximate SNR range 0~25 dB. The improvement of SEEVOC over LP is about 1 dB over much of this range. However, if the SNR is larger than about 30 dB, LP analysis performs slightly better than SEEVOC. It is concluded that, in the important low SNR range, the SEEVOC algorithm is more immune to white Gaussian noise than LP analysis for spectral envelope estimation of voiced speech.

6. CONCLUSIONS

In this paper, we evaluated the performance of the SEEVOC algorithm on synthetic speech with exactly known spectral envelopes, using spectral distortion between the true envelope and the SEEVOC envelope as a performance criterion. The results give new insight into the algorithm, which is very important for low bit rate coding of speech based on the sinusoidal representation.

Theories were developed that can effectively optimize the SEEVOC algorithm and reduce the effects of inaccurate choice of the coarse pitch *CP*. We found that the it is in fact important to start with a reasonably accurate value of coarse pitch *CP*.

This analysis was extended to find the optimum search range parameters α and β, depending on the characteristics of the coarse pitch estimator and on the noise level. Considerable improvements in accuracy as well as computational savings can be obtained in some cases.

It was also found that the SEEVOC algorithm outperforms LP analysis in the presence of noise the SNR range 0~25 dB, and gives similar results at higher SNRs.

REFERENCES

[1] D.B. Paul, "The spectral envelope estimation vocoder", *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29, 1981, pp. 786-794.

[2] R.J. McAulay and T.F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation", *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34, 1986, pp. 744-754.

[3] R.J. McAulay and T.F. Quatieri, "Low bit rate speech coding based on a sinusoidal model", in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi (Eds.), Marcel Dekker, New York, pp. 165-208, 1992.

[4] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.