



Morphological pre-processing technique and its applications on speech signal

Hyun Soo Kim

Telecommunication R&D Center, Samsung Electronics, Korea

Abstract

The properties and applications of morphological filters for speech analysis are investigated. We introduce and investigate a novel nonlinear spectral envelope estimation method based on morphological operations, which is found to be very robust against noise. This method is also compared with the spectral envelope estimation vocoder (SEEVOC) method. A simple method for the optimum selection of the structuring set size without using pitch information is proposed. Also, a new concept of higher order peaks is introduced and found to be beneficial. The morphological approach is then used for a new pitch estimation method. The harmonic-plus-noise decomposition is used to develop a novel and flexible noise reduction method.

Index Terms: harmonic peak, Spectrum, Pitch

1. Introduction

All morphological operations depend on the concept of fitting a structuring element [2]. The structuring set is a sort of sliding window which is symmetric about the origin, and it determines the performance of the morphological operation. The length of the window is twice the size of the structuring set (SSS) plus one; i.e. $L_{window} = 2 * SSS + 1$.

In this paper we mainly investigate morphology in the context of spectral envelope estimation of speech, where the morphological operators are used to extract features of speech signals and to estimate spectral envelopes by selecting true harmonic peaks in the signal. Dilation has proved to be a successful method for spectral estimation and different methods for spectral envelope estimation based on it are proposed here.

The closing operation in the frequency domain has also been found useful for other purposes. Closing is defined as dilation followed by erosion, and tends to remove valleys in the spectrum, leaving the harmonic regions behind. The resulting signal is thus useful for various speech analysis purposes, including pitch determination algorithms (PDAs), harmonic-plus-noise decomposition (HND) and noise reduction. We present here a new pitch determination algorithm based on the harmonic product (or sum) spectrum method [6]. Extensions of our approach to HND are also proposed and investigated.

2. Spectral estimation using morphology

Dilation, one of the fundamental morphological operations, has proved to be a successful method for spectral envelope estimation. We use it to pick major harmonic peaks, and then interpolate (using cubic splines) to obtain the spectral envelope.

When applied to one-dimensional signals, the dilation operator is equivalent to the simple concept of finding the maximum value under a sliding window. It leads to flattening of the spectrum near the peaks. The dilated region increases as the SSS increases, and selecting the optimum SSS for

spectral estimation is closely related to (coarse) pitch estimation.

With large sliding windows (i.e. large SSS values), neighbouring peaks may affect the length of the dilated region. This interference will be less for small sliding windows. It was observed that the choice of SSS affects the performance of spectral estimation.

Three different methods for spectral envelope estimation using dilation were proposed and investigated. In the first method (hitting peak method), we focused on the point where a peak touches the dilated region. When a peak touches the flattened region of the dilated spectrum, the peak is selected and the spectral envelope is estimated by interpolation of the chosen peaks. This strategy proved to be very simple and useful but the performance depends very much on the selection of SSS .

In the second method (mid point method) we picked the mid point of each dilated region. The mid points of all the dilated regions are selected and interpolated to estimate the spectral envelope of the signal. In the hitting peak method, we selected all those points where peaks hit the dilated regions but in mid point method, we only select the mid point of each dilated region. Therefore the chance of selecting low-level peaks is reduced for small SSS . If several peaks lie under a dilated region, only the highest amplitude peak will be selected and others will be ignored in the first method. In the second method, the mid point of each dilated region was picked up and we get reasonably smooth spectra for large SSS . This proved to be better than the first method for small and large SSS , and is relatively tolerant in the selection of low level noisy peaks.

The best of these is the "tracking peak" method, in which, for each dilated region, we select the spectral peak that causes that region to be dilated (it is the largest peak under all sliding windows that include at least part of the dilated region). This strategy selects most true harmonic peaks for interpolation.

This method was compared with the SEEVOC algorithm, using test signals with known spectral envelopes. To obtain such test signals, synthetic voiced speech signals were used, generated by applying trains of impulses to vocal tract filters derived from actual speech signals by linear prediction analysis. Thus the frequency response of the LP vocal tract filter is the true envelope of the synthetic speech signal. This allows exact comparison of any spectral envelope estimation method with the true envelope.

The spectral distortion SD between the true envelope and an estimated envelope was used as the objective measure. This is calculated by the formula

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

where L is the number of frequency points, PX_i is the power spectrum at the i^{th} frequency (true envelope) and PY_i is the corresponding estimated envelope. The gain factor K , which is included to allow for the fact that the scales of the two spectra may be different, is chosen to minimize SD .

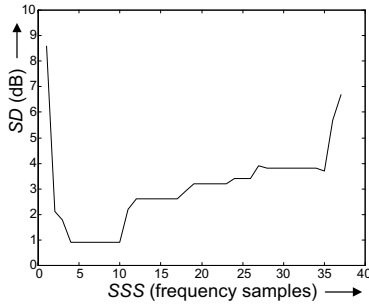


Figure 1: *The effect of SSS on the estimated spectral envelope distortion SD using morphology.*

Figure 1 shows the effect of SSS on the estimated envelope using morphology (dilation, peak picking and interpolation). The true pitch (TP) is 12.8 samples in this case. When SSS is very small, the dilated spectrum follows the shape of the original signal, which leads to very large SD . As SSS starts to increase, some low level peaks are still selected along with all of the harmonic peaks. The performance of the algorithm improves with increasing SSS. The minimum SD is achieved for optimum SSS ranges in which only the major harmonic peaks are selected. In this example, the minimum SD is about 1 dB, which is typical, and this is obtained when the sliding window size is in the range 9–21 samples (i.e. $SSS = 4 - 10$). SD rises steadily but gracefully (with jumps at one pitch period interval in most cases) as SSS increases further, due to missing harmonic peaks. Hence it is important that the SSS value is not substantially underestimated, whereas the morphological method is relatively tolerant of overestimated SSS values.

The analogous effect to the choice of SSS is the choice of coarse pitch CP in the SEEVOC method. [4] It was found that the minimum SD was obtained when the CP is in the range 10 - 17 samples using the same synthetic voiced signal as in this paper. Since the relative range of SSS for optimum performance is greater than that of CP , it follows that the choice of SSS in the new method is less sensitive than the choice of CP in the SEEVOC method.

3. Spectral estimation using morphology in the presence of noise

The morphological spectral estimation method described above is a nonlinear peak selection method that tries to select only the harmonic peaks. It 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 effect of the choice of SSS on the morphological methods in a noisy environment. The added noise was white Gaussian, and the input signal-to-noise ratio (SNR) was varied over the range 0–30 dB. The SD curves in a typical case are shown in Figure 2 (solid line). For SNR above about 25 dB the performance is nearly as good as for infinite SNR . But at low $SNRs$ the SD deteriorates, as expected, although the choice of SSS becomes less critical at low $SNRs$. That is, the range of optimum SSS is larger at low $SNRs$.

We have further improved the performance of this method by making use of a novel *higher order peak* concept. If we call the peaks of the spectrum the first order peaks, then the second order peaks are defined as the peaks of the series (in

the frequency domain) formed by the first order peaks – that is, they are the “peaks of the peaks”. Third and higher order peaks can be defined in a similar way. (The same concept can be used in the time domain, but this is not considered here.)

The nature of the higher order peaks is to have higher levels, on average, than lower order peaks. Higher order peaks also occur less often – there are fewer second and third order peaks than first order peaks. Peak rate of occurrence is an interesting feature of higher order peaks, and it is the second and third order peaks that contain more reliable pitch period information. Another interesting peak feature that can be made use of is the number of points between second or third order peaks. It should be noted that most true harmonic peaks will be among the second or higher order peaks.

The major reason for large SD with small values of SSS is that some low level peaks are chosen for the spectral envelope estimation. Use of the second (or higher) order peaks for the interpolation step gets rid of most of these spurious noisy peaks, thus reducing SD with small values of SSS. For larger SSS values, missing harmonic peaks increase the SD and using higher order peaks for interpolation does not affect the SD . The improvements when using second order peaks for interpolation are shown in Figure 2 (dashed lines), where the results of using first order peaks are also shown (solid lines).

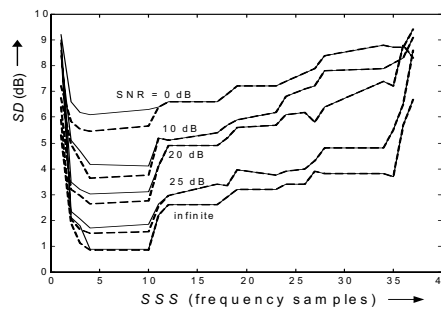


Figure 2: *Performance of the morphological method at different SNRs, showing the effect of the choice of SSS on spectral distortion SD. Solid lines: 1st order peaks used; Dashed lines: 2nd order peaks used.*

4. Speech analysis using morphological pre-processing and optimum SSS selection

Morphological filtering is very simple to implement and it can be used to extract peak or valley features from arbitrary signals. In addition, it is nonparametric; i.e. it does not presuppose anything about the speech spectrum. Some further applications will now be discussed.

The closing operation fills the valleys in a spectrum, and this behaviour can be used for pitch estimation and other operations. It was observed that harmonic peaks stand out and small peaks are flattened after closing with the optimum SSS. With larger SSS, the flattened regions created in the valleys lie above the low peaks, and this can be used for formant estimation.

We consider a pre-processor for the further analysis of speech that consists of the closing operation on the spectrum, followed by the generation of a remainder (or residual) spectrum which is essentially the spectrum components that stand above the closure floor, as shown schematically in Figure 3b. This remainder spectrum is very useful for many subsequent speech analysis techniques. It is also very robust against noise, since it selects the peaks of the harmonics, which stand out even among noise.

If SSS is too low there are too many noise peaks, but if it is too large we miss harmonics. A simple adaptive strategy can find the optimum value, as follows. From the remainder spectrum (Figure 4b), we calculate P , which is the ratio of the energy in the N largest peaks to the total energy in the remainder spectrum, where N is a small number (e.g. 5 - 15) that is also adaptively chosen. The criterion we will use is that the optimum choice of SSS will result in a particular value of P (e.g. 0.3 - 0.5). The optimum choices of N , P and SSS depend on the pitch and SNR , but are not critical. For example, N should be smaller for female than for male speakers, since there will be fewer harmonics in total.

To test this concept we plot the energy ratio P against the normalized structuring set size SSS/TP for many voiced frames. These traces include a variety of pitches TP , N values and $SNRs$ (with N determined by the iterative procedure described in the following). The energy ratio is low up to about $SSS/TP = 0.5$ but there is a very sharp increase at larger values, due to the remainder spectrum missing true harmonic peaks. Therefore, our proposed strategy for SSS selection is to choose the SSS value where the energy ratio suddenly rises, e.g. where P is about 0.3 - 0.5. This results in a sliding window length close to one pitch period.

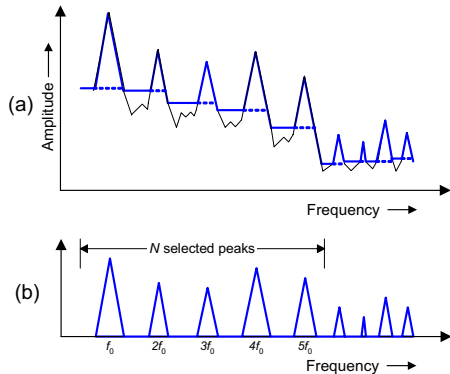


Figure 3: (a) Noisy spectrum and its closure; (b) Remainder or pre-processed spectrum (the part of the spectrum above the closure floor).

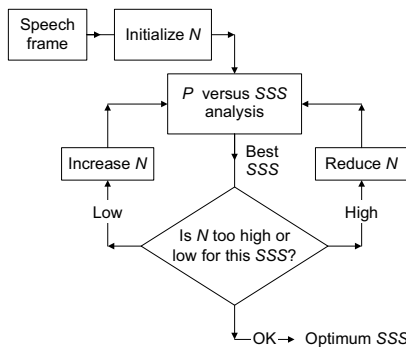


Figure 4: Flow diagram for (automatic and iterative) optimum selection of SSS in the morphological analysis system.

The flow diagram for the resulting adaptive choice of the parameters is shown in Figure 4. We have found that a good choice of N is 7 for high pitch voiced frames, 10 for mid pitch and 13 for low pitch. In this algorithm we start with $N=10$ and then find a value of SSS that gives P in the range 0.3 - 0.5. If this value is too large for the assumed N , we reduce N and do the analysis again to get the final optimum SSS . We do the opposite if SSS is too small.

The final pre-processed signal (remainder spectrum) is extracted using the optimum SSS value. Note that we do not assume any prior pitch information in this algorithm.

5. PDA using morphological preprocessing

We have developed a pitch determination algorithm (M-PDA) using the foregoing morphological pre-processing. It is based on the fact that most of remainder peaks after morphological pre-processing are due to major sine wave components, so that pitch is emphasized for harmonic signals.

The M-PDA is basically morphological pre-processing followed by the harmonic sum spectrum (or harmonic product spectrum) PDA method [6], illustrated in Figure 5. Sinusoidal components at the fundamental frequency f_0 tend to dominate the summed spectrum in Figure 6d. It was found that use of frequency scale compression factors up to five results in good pitch estimates (higher values are even better).

This method works well even when the pre-processed signal misses some of the harmonics, since most of the peaks of the remainder signal are at multiples of the pitch frequency.

The most impressive aspect of the M-PDA is that it works for any SSS . For very small SSS it is equivalent to the original Schroeder method. For the ideal SSS , the pre-processed signal has mainly harmonic peaks, thus emphasizing the fundamental frequency in the summed spectrum. For very large SSS , some of the harmonic peaks will be missing but very strong harmonic peaks will remain, thus still strongly emphasizing the correct pitch frequency. It is easy to appreciate that this method will be largely unaffected by the formant structure. In summary, the M-PDA is very simple and efficient, and does not require any prior pitch information or any assumption about the signal (apart from near harmonicity for voiced sounds).

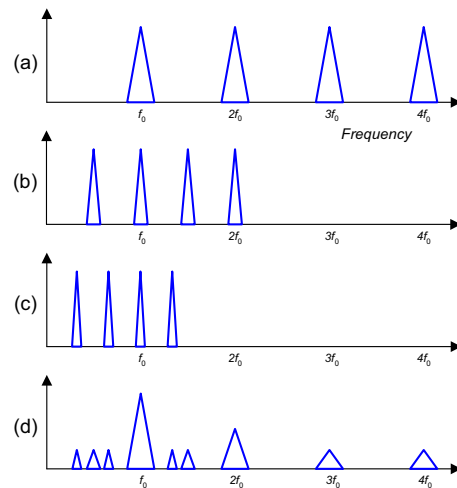


Figure 5: PDA based on the harmonic sum (or product) method with morphological pre-processing. (a) Remainder spectrum; (b) and (c) Frequency scale divided by 2 and 3, respectively; (d) Sum of a - c (scaled).

The M-PDA was tested against two well known PDAs using 300 real voiced speech frames in white noise at various $SNRs$. The others are the autocorrelation (A-) PDA [5] and the sinusoidal (S-) PDA, which is based on a sinusoidal model of speech and which is used in sinusoidal coding [3].

The gross error results (pitch halvings or doublings) are tabulated in Table 1, and the fine error results (no pitch halvings or doublings) are given in Table 2. It is obvious from the results that the A-PDA suffers from severe gross

error problems, due to its inability to distinguish the strong influence of formant structures. Even with the aid of centre-clipping this effect could not be reduced much. The S-PDA is known to be relatively immune to gross pitch errors, but it still suffered from some noise errors and pitch doublings and halvings because of its inability to cope with rapidly rising or decaying regions such as the transition from unvoiced to voiced speech. However, the M-PDA is better than the S-PDA in regard to both gross and fine errors.

This PDA method can be generalized to many other PDA methods (some of which are better than the Schroeder method used), since any existing or proposed PDA can be applied to the morphologically pre-processed signal instead of the original speech signal, with improvements to be expected because of the increased harmonic content of the pre-processed signal.

Error type (SNR)	M-PDA	S-PDA	A-PDA
PD (Clean)	0 (0%)	8 (2.7%)	11 (3.7%)
PH (Clean)	0 (0%)	3 (1%)	15 (5%)
PD (20 dB)	3 (1%)	11 (3.7%)	17 (5.7%)
PH (20 dB)	2 (0.7%)	4 (1.3%)	19 (6.3%)
PD (10 dB)	9 (3%)	15 (5%)	28 (9.3%)
PH (10 dB)	5 (1.7%)	8 (2.7%)	34 (11%)
PD (0 dB)	19 (6.3%)	28 (9.3%)	47 (15.7%)
PH (0 dB)	17 (5.7%)	19 (6.3%)	44 (14.7%)

Table 1: Gross pitch error results. These are the number of frames with pitch halving (PH) or doubling (PD), also shown as a percentage of all frames.

SNR	M-PDA	S-PDA	A-PDA
Clean	0.53%	0.65%	1.06%
20 dB	0.6%	0.84%	1.13%
10 dB	1.28%	1.38%	1.74%
0 dB	2.1%	2.14%	3.82%

Table 2: The standard deviation of the relative pitch error $(TP - EP)/EP$ as a percentage, where TP is the true pitch and EP is the estimated pitch (gross error frames not included)

6. Extension of harmonic-plus-noise decomposition to sinusoidal analysis

The HND method [1] assumes that the input speech signal is harmonic and that the pitch information is available. An obvious change in these methods is to use M-PDA instead of the PDA used in the original HND methods.

However, and more significantly, we can extend the HND method to include the non-harmonic, or general sinusoidal, case by replacing one of the initial steps in the HND (where we identify the sinusoidal regions) with the morphological analysis method. Simple peak picking of the remainder spectrum with the optimum SSS (or larger) will give an accurate initial estimate of the sinusoidal regions. This works as well in the general case as in the harmonic case. It gives an alternative to the methods used in the sinusoidal analysis method [3].

This is illustrated in Figure 6. The search regions A and B that are used in the iterative HND methods are shown in Figure 8a for the harmonic case, and in Figure 8b for the general sinusoidal case. (Regions A surround presumed

sinusoidal components, whereas regions B contain presumed noise or window artefacts.)

This approach may be further extended to include a variable number of sine components (e.g. dependent on the strengths of the various sine waves), or to cases where the A and B regions are of varying width. It also allows the frequencies of the individual components to be accurately estimated.

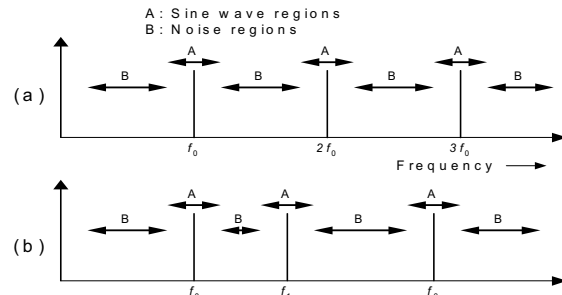


Figure 6: Concept diagram of the extension of the harmonic-plus-noise decomposition method to the more general sinusoidal case. (a) Harmonic case; (b) General sinusoidal case with $f_1 \neq 2f_0$, $f_2 \neq 3f_0$.

7. Conclusion

The properties and some applications of morphology to speech analysis have been investigated. We first introduced and investigated a novel morphological method for spectral envelope estimation. The structuring set size SSS in the morphological method plays the same role as the coarse pitch CP in the SEEVOC method, although it is less critical and can easily be found using a simple iterative procedure. A novel concept of higher order peaks was found to be beneficial. The new method is very robust against noise.

A new robust pitch estimation method (M-PDA) based on the morphologically pre-processed spectrum was also proposed and compared with two other PDAs.

Finally, we showed how our approach may be used to extend the HND method to the general sinusoidal analysis of speech or audio signals.

8. References

- [1] d'Alessandro, C., Yegnanarayana, B. and Darsinos, V. "Decomposition of speech signals into deterministic and stochastic component". Proc. ICASSP: 760-763., 1995
- [2] Maragos, P, Hanson, H.M. and Potamianos, "A system for finding speech formants and modulations via energy separation", Speech and Audio Processing, IEEE Transactions on Volume 2, Issue 3: 436 – 443., 1994
- [3] McAulay, R.J. and Quatieri, T.F., "Pitch estimation and voicing detection based on a sinusoidal speech model". Proc. ICASSP: 249-252., 1990
- [4] Paul, D.B., "The spectral envelope estimation vocoder". IEEE Trans. Acoust., Speech and Signal Proc., vol. ASSP-29: 786-794., 1981
- [5] Rabiner, L.R., "On the use of autocorrelation analysis for pitch detection". IEEE Trans. Acoust., Speech and Signal Proc., vol. ASSP-25: 24-33., 1971
- [6] Schroeder, M.R., "Period histogram and product spectrum: New method for fundamental-frequency measurement". J. Acoust. Soc. Am., vol. 43, no. 4: 829-834., 1968