Improving Pitch Tracking Performance in Hard Noise Conditions by a Preprocessing Based on Mathematical Morphology

Pere Martí-Puig, Jordi Solé-Casals, Ramon Reig-Bolaño

Digital Technologies Group, University of Vic, Sagrada Família 7,
08500 Vic, Spain
{pere.marti, jordi.sole, ramon.reig}@uvic.cat

Abstract. In this paper we show how a nonlinear preprocessing of speech signal -with high noise- based on morphological filters improves the performance of robust algorithms for pitch tracking (RAPT). This result happens for a very simple morphological filter. More sophisticated ones could even improve such results. Mathematical morphology is widely used in image processing and has a great amount of applications. Almost all its formulations derived in the two-dimensional framework are easily reformulated to be adapted to one-dimensional context.

Keywords: Robust Pitch Tracking, Mathematical Morphology, Nonlinear Speech Preprocessing

1 Introduction

Pitch is a very important parameter in speech processing applications, such as speech analyzing, coding, recognition or speaker verification. Pith tracking also becomes relevant for the automatic recognition of emotions in spoken dialogues. Affective activity causes physiological variations reflected in the vocal mechanism and causes further speech variation being the pitch the most relevant acoustic parameter for the detection of emotions (Mozziconacci and Hermes, 1998, Juang and Furui, 2000, Petrushin, 2000 and Kang et al., 2000) [1-4]. For example, aroused emotions (such as fright and elation) are correlated with relatively high pitch, while relaxed emotions (such as tedium and sorrow) are correlated with relatively low pitch.

Pitch detection techniques are of interest whenever a single quasi-periodic sound source is to be studied or modeled [5][6]. Pitch detection algorithms can be divided into methods which operate in the time domain, frequency domain, or both. One group of pitch detection methods uses the detection and timing of some time domain feature. Other time domain methods use autocorrelation functions or some kind of difference of norms to detect similarity between the waveform and its time delayed version. Another family of methods operates in the frequency domain with the purpose of locating peaks. Other methods use combinations of time and frequency domain techniques to detect pitch. Frequency domain methods need the signal to be frequency transformed, and then the frequency domain representation is inspected for the first harmonic, the greatest common divisor of all harmonics. Windowing of the signal is recommended to avoid spectral spreading, and depending on the type of window, a minimum number of periods of the signal must be analyzed to enable accurate location of harmonic peaks [5] [6]. Various linear preprocessing steps can be used to make the process of locating frequency domain features easier, such as performing linear prediction on the signal and using the residual signal for pitch detection. Performing nonlinear operations such as peak limiting also simplifies the location of harmonics. Although there are many methods of pitch estimation and tracking, both in time and frequency domains, accurate and robust detection and tracking is still a difficult problem. Most of theses methods are based on the assumption that speech signal is stationary in short time, but speech signal is non-stationary and quasi-periodical. Among these methods, autocorrelation-based method is comparatively robust against noises, but it may result in a half-pitch or double pitch error, and if noise is high, this method can’t detect pitch properly. In this paper we improve the performance of robust algorithms for pitch tracking (RAPT) by means of a nonlinear preprocessing whit a filter based on mathematical morphology. The used RAPT is due to D. Talkin [7, 8] with only two minor differences.

2 The Mathematical Morphology

Mathematical morphology was proposed by J.Serra and G. Matheron in 1966, was theorized in the mid-seventies and matured from the beginning of 80’s. Mathematical morphology is based on two fundamental operators: dilation and erosion. It can process binary signals and graylevel signals and it has found its maximum expression in image processing applications. These two basic operations are done by means of a structuring element. The structuring element is a set in the Euclidean space and it can takes different shapes as circles, squares, or lines. Using different structuring elements it will achieve different results; therefore, the election of an appropriate structuring element is essential. A binary signal can be considered a set and dilation and erosion are Minkowski addition and subtraction with the structuring element [9]. In the context of speech processing we work with graylevel signals. In this context, the addition and subtraction operations in binary morphology are replaced by suprermum and infimum operations. Moreover, on the digital signal processing framework, supremum and infimum can be changed by maximum and minimum operations.

We define the erosion as the minimum value of the part of the image function in the mobile window defined by the structuring element, Y, when its beginning is situated on *x* (one-dimensional framework) or in *x,y* (two-dimensional framework). As we deal with speech signals we are interested in one-dimensional definitions. Then, given the one-dimensional signal, the *f* function, and the flat structuring element, Y, the erosion can be defined as:

$$\varepsilon_Y(f)(x) = \min_{s \in Y} f(x+s) \tag{1}$$

The erosion uses the structuring element as a template, and gives the minimum graylevel value of the window function defined by the mobile template; decreasing peaks and accentuating valleys (see fig.1 b).

On the other hand the graylevel signals dilation is defined as:

$$\delta_Y(f)(x) = \max_{s \in Y} f(x-s) \qquad \textbf{(2)}$$

The dilation gives the maximum graylevel value of the part of the function included inside the mobile template defined by the structuring element, accentuating peaks and minimizing valleys. By combining dilation and erosion we can form other morphological operations. Opening and closing are basic morphological filters.

The morphological opening of a signal *f* by the structuring element Y is denoted by $\gamma_Y(f)$ and is defined as the erosion of *f* by Y followed of dilation by the same structuring element Y. This is:

$$\gamma_Y(f) = \delta_Y(\varepsilon_Y(f)) \qquad \textbf{(3)}$$

And the morphological closing of a signal *f* by the structuring element Y is denoted by $\varphi_Y(f)$ and it is defined as the dilation of *f* by Y followed of the erosion by the same structuring element:

$$\varphi_Y(f) = \varepsilon_Y(\delta_Y(f)) \qquad \textbf{(4)}$$

Opening and closing are dual operators. Closing is an extensive transform and opening is an anti-extensive transform. Both operations keep the ordering relation between two images (or functions) and are idempotent transforms. [9]. In the image context the morphological opening removes small objects from an image while preserving the shape and size of larger objects, and the morphological closing fills the gaps between objects. In the one-dimensional context both operations -by means of a non-linear process- create a more simple function than the original.

By combining an opening and a closing, both of them with the same structuring element, we can only create four different morphological filters. Then, considering the operations γ_Y and φ_Y, the four filters we could obtain are $\gamma_Y\varphi_Y$, $\varphi_Y\gamma_Y$, $\gamma_Y\varphi_Y\gamma_Y$ and $\varphi_Y\gamma_Y\varphi_Y$. From the composition of γ_Y and φ_Y no other different filter can be produced as a consequence of idempotency property.

To derive different families of morphological filters we need to combine openings and closings whit different structuring elements. There is a well known method to obtain new filters by alternating appropriately theses operators. The resulting filters are called alternating sequential filters [9] which are very effective tools to fight against noise.

In Fig.1 we can see how these morphological operators work. In Fig.1(a) we have represented a fragment of 0.1ms of speech signal, in (b) it appears the original signal (in black), a dilation (in red) and an erosion (in blue), and in (c) there are the original signal (in black), a morphological close (in red) and a morphological open (in blue). All the morphological operators involved in fig.1 use a flat structuring element of length 60 samples (3.75ms).

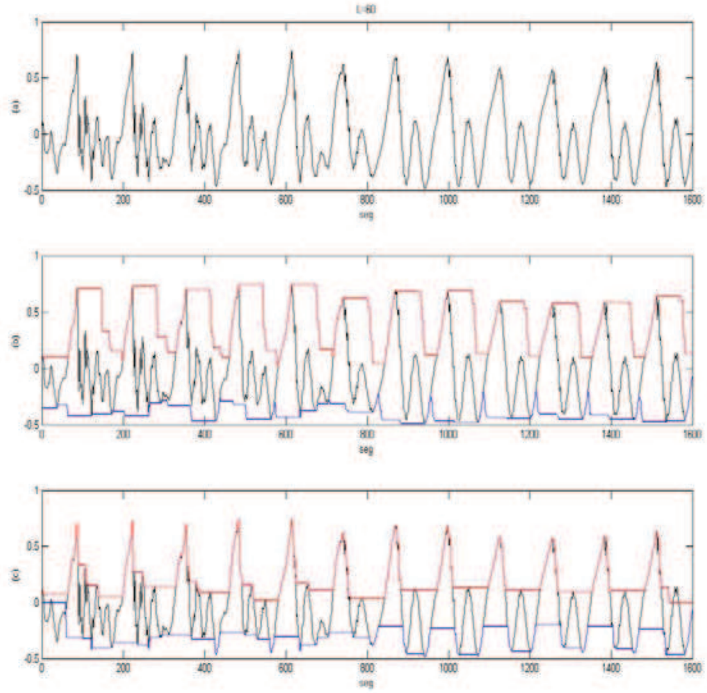

Fig. 1. (a) Original signal (b) red: dilation, blue: erosion (c) red: morphological closing, blue: morphological opening. All results obtained using a flat structuring element of L=60 (3.75ms).

In fig.2 we apply a closing with different structuring elements on the same input signal. We can appreciate the variation of the results depending on the length of the structuring element.

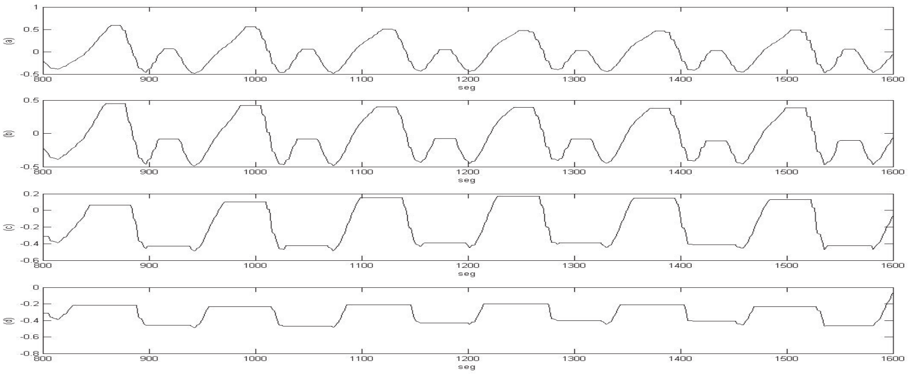

Fig. 2. Morphological closings -by structuring element length L -of the same signal (a) L=10 (b) L=20 (b) L=40 (b) L=60.

3 The Selected Pitch Tracking Algorithm

In order to track the pitch we have used free software provided in the voicetoolbox for MATLAB that can be modified and redistributed under the terms of the GNU - General Public License- [8]. The Robust Algorithm for Pitch Tracking (RAPT) is taken from the work of D. Talkin [7] with only two differences. The first is related whit the modification of the coefficient AFACT which in the Talkin algorithm corresponds approximately to the absolute level of harmonic noise in the correlation window. In the used version this value is calculated as the maximum of three figures: (i) an absolute floor set, (ii) a multiple of the peak signal and (iii) a multiple of the noise floor [8]. The second difference is that the LPC used in calculating the Itakura distance uses a Hamming window rather than a Hanning window.

The software plots a graph showing lag candidates of pitch values and draws the selected path. This original signal representation could be seen in fig.4 where in the upper side there, in blue, the parts detected as voice and in red the parts detected as silent. Down, with red crosses indicating the beginning of a frame, there is represented the possible pitch values and the evolution of the selected path are depicted with continuous blue line. The pitch is given in time units (period). This pitch tracking algorithm is very robust and maintains a good performance under hard noise conditions. However, as the signal-to-noise ration increases the estimation falls into errors. To show these limitations we have introduced an additive white Gaussian noise to the same fragment of signal represented in fig. 4. In next fig. 5 we represent its behavior under three different conditions. The additive white Gaussian noise introduced in the signal has an effect on the entire voice band. In order to improve the performance of the RAPT we propose a nonlinear preprocessing filtering base on the mathematical morphology. In the left part of fig. 5 we can see the RAPT performance when the input signal has a SNR of 0,5dB the graphic shows that only some parts of the original speech are recognized as voice. In the right hand side of fig. 5 we can see the RAPT performance when the input signal has a SNR of -3.5dB; in those conditions the algorithm doesn't work.

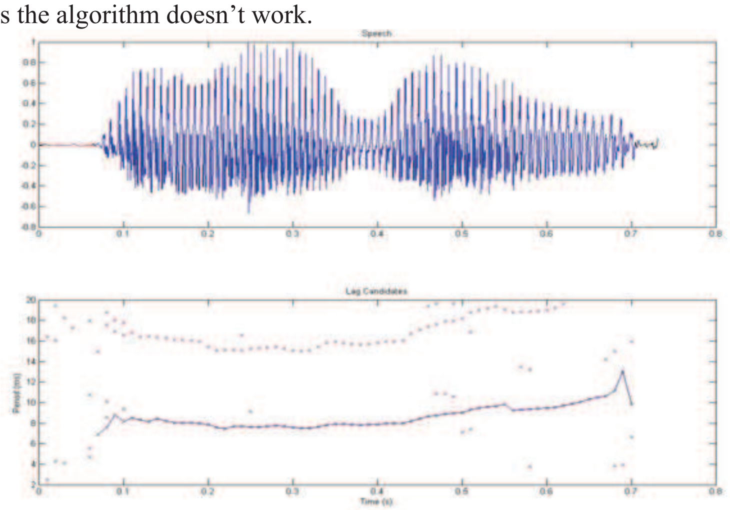

Fig. 4. Pitch evolution given by the RAPT.

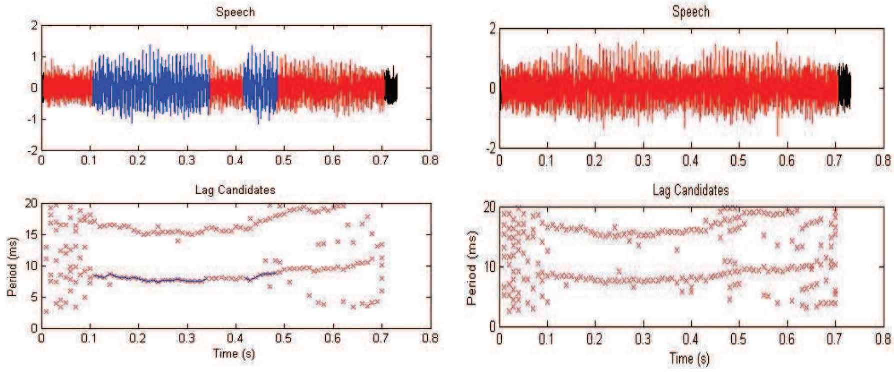

Fig. 5. Left: RAPT performance when the input signal has a SNR of 0.5 dB; only some parts of the original speech are recognized as voice. Right: RAPT performance when the input has a SNR of -3.5 dB; in those conditions the algorithm doesn't work.

4 Signal Preprocessing Based on Mathematical Morphology

In order to obtain a new representation of the noisily input signal that preserves the pitch information we propose the application of morphologic filters in a preprocessing stage. In this work we deal only with flat structuring elements. To design the appropriate filter we have explored different morphologic filters configurations with different structuring element lengths. Those studies had been done using a speech database. We have found that the input signal preprocessing by very simple filters like the compositions $\varphi_5\gamma_5$ or $\varphi_3\gamma_3$ improves the RAPT performance. Theses results could be appreciated in fig. 6 for the same fragment of the signal represented in fig. 4. In fig. 6 the signal is corrupted with Gaussian noise; in its the left side we have applied this signal directly to the RAPT algorithm and in the right side we have a morphological preprocessing by $\varphi_3\gamma_3$.

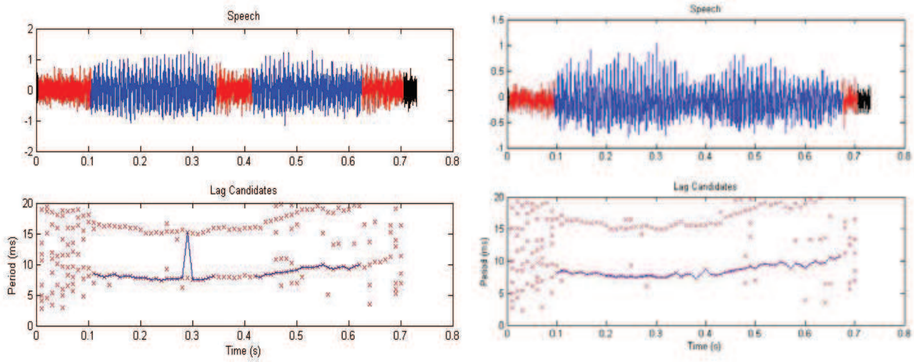

Fig. 6. RAPT performance. (Left) Input signal of SNR of 0.5 dB. (Right) The input signal of SNR of 0.5 dB had been previously preprocessed by $\varphi_3\gamma_3$.

More the sophisticated filters improve theses results. We propose $\varphi_4\gamma_4\varphi_3\gamma_3\varphi_2\gamma_2$.

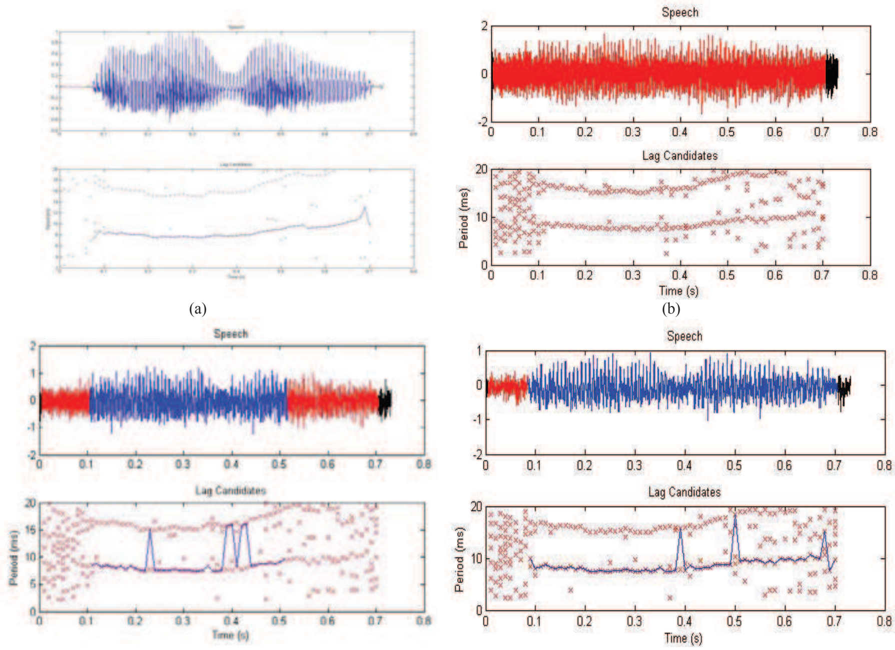

Fig. 6. RAPT results for an input: (a) without noise (b) SNR=-3.5 dB (c) SNR = -3.5dB preprocessing by $\varphi_3\gamma_3$ (d) SNR=-3.5dB and preprocessing by $\varphi_4\gamma_4\varphi_3\gamma_3\varphi_2\gamma_2$.

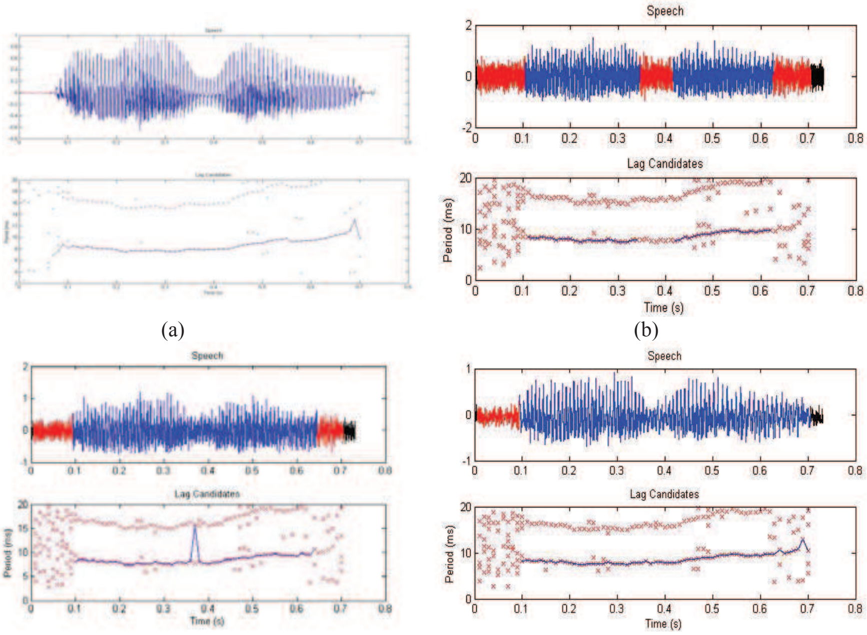

Fig. 7. Same results than fig. 6 changing the SNR to -0.5dB.

6 Conclusions

In this paper we have shown how a pre-processing based on mathematical morphological filters improves the performance of pitch trackers when the input signal are corrupted with additive white Gaussian noise. The nonlinear mathematical morphology techniques are widely developed in image processing and its results can often been exported to the one-dimensional framework. From our knowledge those techniques are not widely explored in speech processing. In [10] we have found a work that also uses simple morphologic filters to estimate the pitch. The objective of [10] is quite different of ours: they are interested in the estimator and we are interested in the signal pre-processing. This could be reflected in the design of the morphological filters and in the size of the structuring elements that such filters use.

Acknowledgments. This work has been supported by the University of Vic under the grant R0904.

References

1. Mozziconacci, S., Hermes, D., Study of intonation patterns in speech expressing emotion or attitude: production and perception. IPO Annual Progress Report. IPO, Eindhoven. 1998.
2. Juang and Furui, 2000 B.-H. Juang and S. Furui, Automatic recognition and understanding of spoken language—a first step towards natural human–machine communication, Proc. IEEE 88 (2000) (8), pp. 1142–1165
3. Petrushin, V.A., Emotion recognition in speech signal: experimental study, development, and application. In: Proc. Internat. Conf. on Spoken Language Processing, Beijing, China. 2000.
4. Kang, B.-S., Han, C.-H., Lee, S.-T., Youn, D.-H., Lee, C., Speaker dependent emotion recognition using speech signals. In: Proc. Internat. Conf. on Spoken Language Processing, Beijing, China, 2000.
5. W. Hess, "Pitch Determination of Speech Signals," Berlin: Springer Verlag, 1983.
6. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399¬418, 1976.
7. D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding & Synthesis, W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.
8. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
9. J. Serra, Image Analysis and Mathematical Morphology. New York: Academic, 1982.
10. Zhao Xiaoqun and Wang Guangyan.: A New Approach of the Morphology Filter for Pitch Contrail Smoothing of Chinese Tone. Signal Processing, 19(4) (2003) 354-357