ROBUST PITCH DETECTION OF SPEECH SIGNALS USING STEERABLE FILTERS

Jinhai Cai cai@cs.mu.OZ.AU

Zhi-Qiang Liu zliu@cs.mu.OZ.AU

Computer Vision and Machine Intelligence Lab
Department of Computer Science
The University of Melbourne
Parkville, VIC 3052, Australia

ABSTRACT

Most of the well known and widely used pitch determination algorithms are frame-based. They only consider the speech local stationarity within the analysis frame. However, our novel pitch determination algorithms employ the steerable filters to obtain the direction of pitch change. Therefore, the proposed algorithms not only make full use of the information within an analysis frame, but also optimally utilize the information from neighbor frames by taking the advantage of the pitch direction. This allows us to use more than one frame to enhance pitch peaks for non-stationary, noisy speech signals. As a result, the proposed algorithms are superior to conventional methods in term of accuracy and reliability, and is robust to noise. Besides, the direction of pitch change can be estimated in different domains. Therefore, our algorithms can be applied in either time or frequency domain, or both of them.

1. INTRODUCTION

Pitch conveys some useful information on speaker and language. Accurate estimation of the pitch period (T_0) or fundamental frequency (F_0) of speech signals is essential to many applications, such as low bit rate speech coding[1], comb filter based speech enhancement[2] and psychiatric researches[3]. For tone languages, pitch is one of the major features used in speech recognition. Therefore, the pitch determination is an important issue in speech processing. Many pitch determination algorithms (PDA) have been proposed in different domains in past decades. A good survey of these algorithms was given by Hess[1]. However, pitch determination still constitutes one of the most problematic topics in speech researches due to the non-stationarity of speech signals and noise corruption. In order to combat the noise corruption, a long analysis frame must be used in conventional algorithms. However, if a significant change of pitch period or formant frequencies occurs over the analysis frame, the smearing effects will be introduced. When a short analysis frame ($\geq 2T_0$) is used, the position of window and the noise corruption will greatly influence the accuracy and reliability of pitch estimation. Many PDAs fail over these non-stationary and low SNR frames. Usually, pitch tracking algorithm[4] is used as the post-processing to reduce the errors in most PDAs. But, some errors are not recoverable.

In this paper, we proposed to use more than one analysis frame to enhance the robustness to noise and to improve the accuracy and reliability of PDAs. Our algorithms are based on the following facts:

- Speech is a dynamic and information-bearing process. Therefore, it remains *stationary* only in short time.
- The rate of pitch period change is a random process in long term. However, it approximates to a constant over few short frames.

Therefore, an optimal filter for pitch detection should have following properties:

- It's a low-pass filter for inter frames in the direction of pitch change. Therefore, it can alleviate the influence of window positions and smearing effect and reduce noise.
- It's a band-pass filter for intraframe to optimally enhance pitch peak.
- Its orientation is tunable to adapt the direction of pitch change.

Two dimensional Gabor filters and the orientational derivative Gaussian filters satisfy these properties. Because the orientational derivative Gaussian filters are steerable and separable[5], they need much less computation than Gabor filters in estimating the direction of pitch change. Therefore, we use the orientational derivative Gaussian filters in our experiments.

2. STEERABLE FILTERS

Oriented filters are widely used in computer vision and image processing, such as motion analysis, edge detection, line parameter estimation and texture analysis. Usually, the motions, edges and lines can be characterized by a set of parameters: position, orientation, velocity and size (or width), etc., and can exist at any possible positions and orientations. In order to adapt their parameters, we should be able to obtain the response of a filter at an arbitrary position and orientation. It is practically impossible to tune the filters to all possible positions and orientations in real-time. Because the computation is too huge by this way. The efficient way is to design a family of filters that any filter in this family can be represented by few basis filters. Therefore, the output of a filter can be expressed as a weighted sum of outputs of the designed basis filters. Such filters are called "steerable filters".

2.1. 2D Steerable Filters

In two-dimensional case, the rotations are operated in the plane. The steerable filters can be expressed as a linear combination of basis functions. The definition of two-dimensional steerable filters[5] is:

$$f^{\theta}(x,y) = \sum_{j=1}^{M} k_j(\theta) g_j(x,y) \tag{1}$$

where $f^{\theta}(x,y)$ represents $f(x,y)$ rotated by an angle θ about the origin, $g_j(x,y)$ is the jth basis filter, $k_j(\theta)$ is the interpolation function of $g_j(x,y)$ and M is the number of basis filters. The minimal number of basis filters required to steer a two-dimensional function depends on the number (N) of harmonics which span the function using a Fourier series in polar coordinates. The basis filter $g_j(x,y)$ is the rotated version of steerable filter $f(x,y)$ at a designed angle.

The steering condition (1) holds if and only if $M \geq$ the minimal number of basis filters required and the interpolation function $k_j(\theta)$ satisfies

$$\begin{pmatrix} 1 \\ e^{i\theta)} \\ \vdots \\ e^{N\theta)} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{i\theta_1)} & e^{i\theta_2)} & \cdots & e^{i\theta_M)} \\ \vdots & \vdots & & \vdots \\ e^{iN\theta_1)} & e^{iN\theta_2)} & \cdots & e^{iN\theta_M)} \end{pmatrix} \begin{pmatrix} k_1(\theta) \\ k_2(\theta) \\ \vdots \\ k_M(\theta) \end{pmatrix}. \tag{2}$$

A convenient way[5] to determining the minimal number of basis filters is to express the steerable filters as polynomials in Cartesian coordinates:

$$f(x,y) = P_T(x,y) W(r), \tag{3}$$

where $W(r)$ is a windowing function, $r = \sqrt{x^2+y^2}$, and $P_T(x,y)$ is a Tth order polynomial in x and y. $T+1$ is the minimal number of basis filters required.

2.2. 2D Separable Steerable Filters

If a steerable filter is with separable bases, we can further greatly reduce the computational costs. The separable filter can be expressed as a polynomial in x and y. Thus, the filter is written as:

$$f(x,y) = \sum_i \sum_j \alpha_{ij}(\theta) x^i y^j W(r), \tag{4}$$

where

$$W(r) = W(x)W(y),$$

and $\alpha_{ij}(\theta)$ are the coefficients. If the kernel size of the steerable filters is $S \times S$, the estimation of orientation using steerable filters becomes an $O(SM)$ process instead of an $O(S^2M)$ process. Therefore, it is important to choose the steerable filters with separable bases if computational costs are vital to the system.

3. THE DETERMINATION OF PITCH CHANGE DIRECTION

According to the theory of speech production, the pitch mainly depends on the glottis, subglottal pressure and tenseness of vocal-cord. Usually, these factors change much slowly compared with the momentary position of the vocal tract. The direction of pitch change also does not vary rapidly. The speech narrow-band spectrum in Fig.1 show that the F_0 and its harmonic peaks appear as smooth curves. Therefore, it is appropriate to use more than one analysis frame to estimate the pitch period.

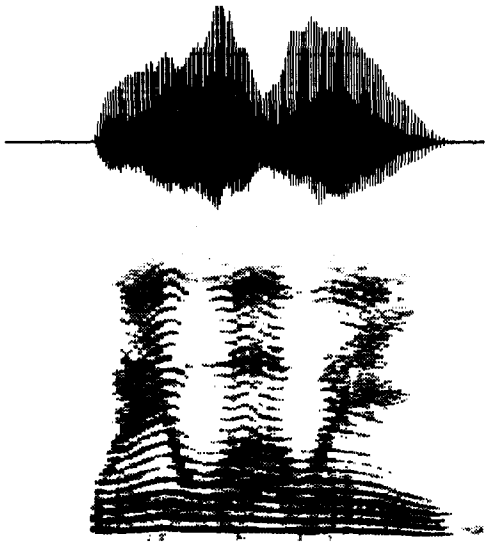

Figure 1. Top: Speech waveform of the sentence "We were away." spoken by a male; Bottom: Spectrum of the speech.

3.1. Orientational Gaussian Filters

Consider the speech features $f(z,t)$ in two dimensional space, where t is time and z is the coordinate in speech processing domain which is time or frequency. The support Gaussian kernel $G(z,t)$ is

$$G(z,t) = k \exp[-(\frac{z^2}{\sigma_z^2} + \frac{t^2}{\sigma_t^2})] \tag{5}$$

where k is a normalization constant and σ_z and σ_t are scaling constants of z and t respectively. We rewrite above equation as

$$G(x,y) = k \exp[-(x^2+y^2)] \tag{6}$$

where $x = z/\sigma_z$ and $y = t/\sigma_t$. The first and second derivative of Gaussian function used in our experiments are:

$$\begin{aligned} G_1(x,y) &= -1.82x \exp[-(x^2+y^2)] \\ G_2(x,y) &= (1-2x^2)\exp[-(x^2+y^2)]. \end{aligned} \tag{7}$$

Let $G_i^{\theta}(x,y)$ represent $G_i(x,y)$ rotated by an angle θ about the origin, where $i = 1$ or 2. Since the rotation does not change the Gaussian function and the polynomial is still a polynomial after rotation operation. According to (1) and (4), it is easy to find that $G_1(x,y)$ and $G_2(x,y)$ are separable and steerable filters. The minimal number of basis filters for $G_1(x,y)$ and $G_2(x,y)$ is 2 and 3, respectively. The chosen basis filters are $G_1^{0^\circ}(x,y)$ and $G_1^{90^\circ}(x,y)$ for G_1, and $G_2^{-45^\circ}(x,y)$, $G_2^{0^\circ}(x,y)$ and $G_2^{45^\circ}(x,y)$ for G_2.

3.2. The estimation of orientation of pitch curves

For pitch detection, two order derivative of a Gaussian is necessary and enough. This is because that there is no junction in pitch curves and higher order derivative of a Gaussian will result in higher computation costs.

Once the bases of steerable filters are determinated, we can apply them to estimate the direction of pitch change. Usually, the orientation is estimated from a Fourier series of oriented energy[5], $E(\theta)$, which is the squared output of $G_2^\theta(x,y)$ and its Hilbert transform. This involves Hilbert transform and Fourier expansion which increase the computation costs. Moreover, the algorithm[5] disregards the higher order terms of the Fourier series. Therefore, it will result in inaccuracy in some cases. In order to avoid these disadvantages, we use $G_1^\theta(x,y)$ to replace the Hilbert transform of G_2^θ and directly use the outputs of $G_i^\theta(x,y)$, which have no higher order terms, instead of the oriented energy.

Let $O_i^\theta(x,y)$ represent the output of oriented filter $G_i^\theta(x,y)$. The outputs of steerable filter G_1^θ and G_2^θ are

$$O_1^\theta(x,y) = cos(\theta)O_1^{0^\circ}(x,y) + sin(\theta)O_1^{90^\circ}(x,y), \quad (8)$$

and

$$\begin{aligned} O_2^\theta(x,y) &= cos(2\theta)O_2^{0^\circ}(x,y) \\ &+sin(\theta)[sin(\theta)+cos(\theta)]O_2^{45^\circ}(x,y) \\ &+sin(\theta)[sin(\theta)-cos(\theta)]O_2^{-45^\circ}(x,y). \end{aligned} \quad (9)$$

Let $\frac{\partial O_2^\theta}{\partial \theta} = 0$, we can obtain two possible orientations ϕ_1 and ϕ_2 which may maximize the oriented energy of $O_2^\theta(x,y)$:

$$\begin{aligned} \phi_1(x,y) &= \frac{\arctan[O_2^{45^\circ} - O_2^{-45^\circ}, 2O_2^{0^\circ} - O_2^{45^\circ} - O_2^{-45^\circ}]}{2}, \\ \phi_2(x,y) &= \frac{\arctan[O_2^{-45^\circ} - O_2^{45^\circ}, O_2^{45^\circ} + O_2^{-45^\circ} - 2O_2^{0^\circ}]}{2}. \end{aligned} \quad (10)$$

The optimal estimation of orientation for $O_2^\theta(x,y)$ is

$$\theta_{d2}(x,y) = \begin{cases} \phi_1(x,y) & \text{if } |O_2^{\phi_1}| > |O_2^{\phi_2}|, \\ \phi_2(x,y) & \text{else.} \end{cases} \quad (11)$$

Similarly, let $\frac{\partial O_1^\theta}{\partial \theta} = 0$, we can obtain the optimal estimation of orientation for $O_1^\theta(x,y)$:

$$\theta_{d1}(x,y) = \arctan[O_1^{90^\circ}, O_1^{0^\circ}]. \quad (12)$$

The final direction of pitch change is determined by the stronger one. It is given by

$$\theta_d(x,y) = \begin{cases} \theta_{d1} & \text{if } |O_1^{\theta_{d1}}| > |O_2^{\theta_{d2}}|, \\ \theta_{d2} & \text{else.} \end{cases} \quad (13)$$

Because, the $G_2^\theta(x,y)$ is a zero phase filter across its center, the peaks of $O_2^{\theta_d}(x,y)$ occur at the same positions of pitch and its its harmonic peaks. Moreover, the orientations at peaks are determined by $O_2^\theta(x,y)$ only. Therefore, $O_2^{\theta_d}(x,y)$ is used for pitch determination.

4. PITCH DETERMINATION

For short-term analysis PDAs, the T_0 is defined as the average pitch duration in a given analysis frame. The proposed PDAs are based on the conventional PADs[6,7] cooperating with steerable filters which are used to enhance the pitch information.

4.1. Pitch detectors based on autocorrelation analysis

The autocorrelation analysis plays an important role in many aspects of speech processing. The short-term autocorrelation function is defined as

$$r_s(\eta;m) = \frac{1}{L}\sum_{n=-\infty}^{\infty} s(n)w(m-n)s(n-|\eta|)w(m-n+|\eta|), \quad (14)$$

where L is the length of analysis frame, $w(n)$ is the window function, η is time lag, $s(n)$ is the speech signal and m is the window position. In our experiments, we use hamming window to alleviate the influence of window position. The conventional algorithm and our proposed method are shown in Fig.2.

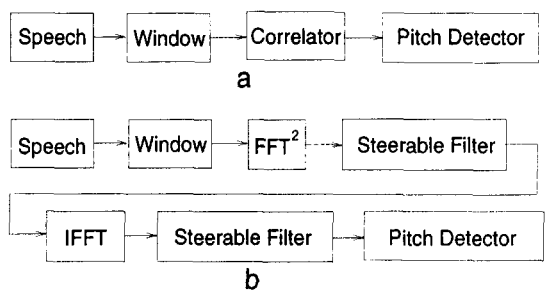

Figure 2. (a) Conventional autocorrelation pitch detector; (b) Oriented filtering-autocorrelation pitch detector.

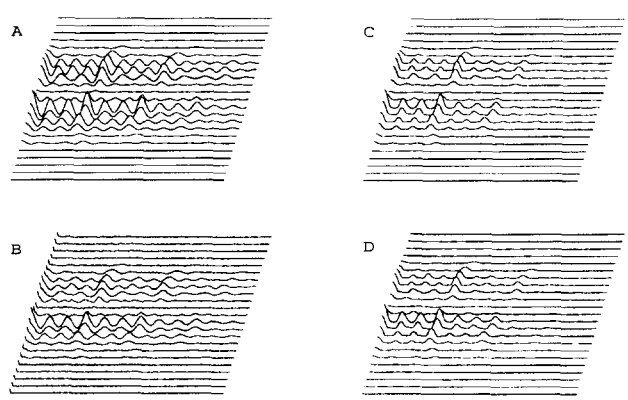

Figure 3. Pitch determination by autocorrelation based algorithms. A, B are the results of conventional algorithm; C, D are the results of proposed algorithm. The speech in A and C is clean, and in B and D is noisy(SNR=0 dB).

In this proposed method, the autocorrelation sequence is computed by FFT. So, the steerable filter can be applied on speech spectrum to enhance the F_0 and its harmonic components. After filtering, the spectrum valleys, which have low energy and are susceptible to noise corruption,

become negative. Therefore, the robustness of the pitch detector to noise can be significantly enhanced by removing the negative parts. The steerable filter applied on autocorrelation sequences further enhances F_0 peaks. Fig.3 shows the results of the two algorithms. Clearly, our proposed algorithm is superior to the conventional one in robustness.

4.2. Pitch detectors based on cepstral analysis

Cepstral analysis offers another way to estimate F_0. Due to the spectral flattening which is performed by taking logarithm, the cepstral analysis based PDAs work well in good acoustic conditions. In comparison with autocorrelation based PDAs, the cepstral PDAs are not susceptible to the influence of speech formants (particularly, the first formant). But, their fatal shortage is that they are very susceptible to noise. Therefore, the robustness to noise is a key issue for this category of PDAs. The steerable filters provide a good solution to the problem of noise corruption. Fig.4 shows the ordinary cepstral PDA and our proposed algorithm.

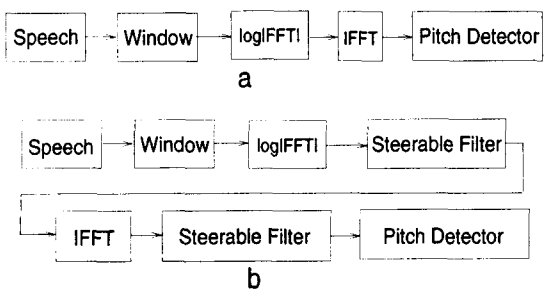

Figure 4. (a) Ordinary cepstral analysis based pitch detector; (b) Oriented filtering-cepstral pitch detector.

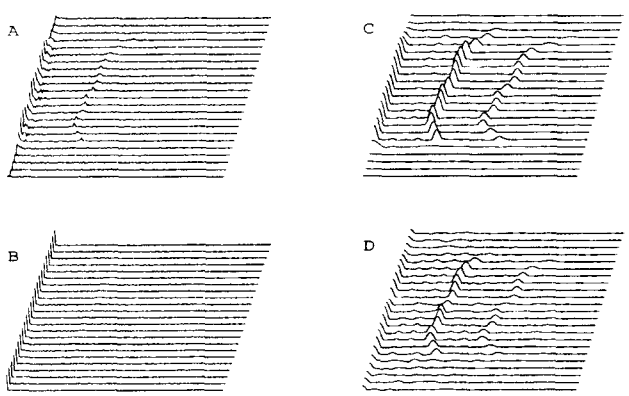

Figure 5. Pitch determination by cepstral analysis based algorithms. A, B are the results of ordinary algorithm; C, D are the results of proposed algorithm. The speech in A and C is clean, and in B and D is noisy(SNR=0 dB).

Similarly, this proposed algorithm applies on the logarithmic spectrum and cepstrum. We also set the negative parts of logarithmic spectrum to zero to reduce noise. Besides, the orientational filter is a low-pass filter along the pitch curves, it further enhances the pitch peaks. This make our proposed algorithm become one of the top robust algorithms to white noise and the results shown in Fig.5 have proven that. Fig.5(B) shows the ordinary algorithm fails at the condition of SNR=0 dB where speech is corrupted by white Gaussian noise. But our proposed one still gives a good result shown in Fig.5(D). Moreover, the low-pass filter along pitch curves takes more information from neighbor frames, it reduces the smearing effects caused by non-stationarity. Another advantage of the filter is that the equivalent frame length is longer than L. Consequently, it greatly alleviates the influence of window positions. The results are shown in Fig.6, where the pitch of clean speech is estimated by two algorithms.

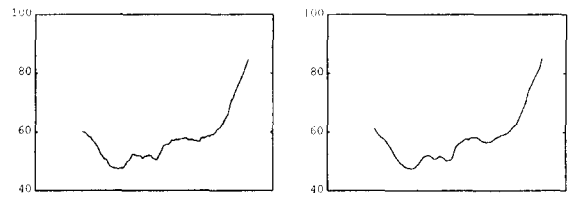

Figure 6. Left: Pitch estimated by ordinary cepstral detector; Right: Pitch estimated by proposed detector.

5. CONCLUSION

In this paper, we have presented two pitch determination algorithms using separable, steerable filters. The oriented filters are applied in both time and frequency domains. Because the filters can adapt the orientations of pitch curves, they optimally utilize the information from neighbor frames. The experimental results show that our proposed algorithms are more robust to white noise, less sensitive to non-stationarity and window positions than conventional algorithms. Besides, the principle of proposed algorithms can be applied on most frame-based PDAs.

REFERENCES

[1] W. Hess, *Pitch determination of speech signals*, Springer-Verlag, 1983.

[2] J.R. Deller, J.G. Proakis and J.H.L. Hansen, *Discrete-time processing of speech signals*, Macmillan Publishing Company, 1993.

[3] Å. Nilsonne,"Acoustic analysis of speech variables during depression and after improvement," ACTA Psychiatrica Scandinavica, Vol.76, pp.235-245, 1987.

[4] B.G. Secrest and G.R. Doddington, "An integrated pitch tracking algorithm for speech systems," proceeding of international Conference on Acoustic, Speech and Signal Processing, pp.1352-1355, 1983.

[5] W.T. Freeman and E.H. Adelson,"The design and use of steerable filters," IEEE trans. on PAMI. Vol.13, pp.891-906, 1991.

[6] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Trans. on ASSP. Vol.26, pp.24-33, 1977.

[7] A.M. Noll, "Cepstrum pitch determination," Journal of the Acoustical Society of America, Vol.14, pp.293-309, 1967.