Robust dynamic tone feature extraction using 2D oriented filters

J. Cai

A novel method is proposed to extract dynamic tone features using 2D oriented filters. As 2D oriented filters are capable of capturing pitch orientations, they can make full use of the information from neighbouring frames to extract dynamic tone features without knowing pitch or using pitch tracking algorithms. Experimental results show that the proposed method can produce dynamic tone features of good quality.

Introduction: Tone features are useful to many speech processing systems [1]. Tone is a main tool for emphasising a syllable, a word, a phrase or even changing the meaning of a sentence. In tonal language, tones carry lexical meanings and play a vital role in distinguishing words with the same pronunciations. Therefore tone features are indispensable to many tonal language speech recognition systems[1, 2].

Tone features are usually derived from fundamental frequency (F_0) or pitch period (T_0) that is extracted from speech by pitch determination algorithms (PDAs) [3]. However, due to irregular glottal excitation and noise corruption, the extracted dynamic features are very noisy. The blame for failing to produce smooth dynamic tone features by existing PDAs should be partially laid on the current criterion of 'good' PDAs. Current PDA criterion focuses only on accuracy of pitch estimation and ignores the smoothness of dynamic features that are sometimes more important.

In this Letter, a new method to extract ΔT_0 without knowing pitch values is proposed. The proposed method is robust to noise and doubled/halved T_0, and can produce smooth normalised ΔT_0.

2D oriented filters for normalised ΔT_0 estimation: As a 2D oriented filter can utilise information from neighbouring frames as well, it is extremely robust to noise. Autocorrelation PDAs are also robust to noise and outperform many computation-intensive PDAs [4]. Therefore it is expected that a combination of these two can work very well at noisy environments.

A. The second derivative of a Gaussian: For dynamic tone feature extraction, it is desirable to have a 2D oriented filter that is a lowpass filter for inter-frames in the direction of pitch orientation. Therefore, it can alleviate the influence of irregular glottal excitations and also enhance the resistance to noise. The second derivative of a Gaussian is a suitable choice according to the above criterion. Let $G_2^0(x, y)$ represent the negative function of the derivative of a Gaussian

$$G_2^0(x, y) = (1 - 2x^2)e^{-(x^2+y^2)} \tag{1}$$

Let $G_2^\theta(x, y)$ be the rotated version of $G_2^0(x, y)$ by an angle θ. $G_2^\theta(x, y)$ can also be expressed as a rotated version of $(1 - 2x^2)$ by an angle θ within the Gaussian envelope $e^{-(x^2+y^2)}$. As $G_2^\theta(x, y)$ is steerable [5], we have

$$\begin{aligned} G_2^\theta(x, y) = {} & k_1(\theta)G_2^{0^\circ}(x, y) \\ & + k_2(\theta)G_2^{45^\circ}(x, y) + k_3(\theta)G_2^{-45^\circ}(x, y) \end{aligned} \tag{2}$$

where interpolation functions $k_i(\theta)$ are given by

$$\begin{aligned} k_1(\theta) &= cos(2\theta) \\ k_2(\theta) &= sin(\theta)[sin(\theta) + cos(\theta)] \\ k_3(\theta) &= sin(\theta)[sin(\theta) - cos(\theta)] \end{aligned} \tag{3}$$

The main advantages of using an oriented filter based on the second derivative of a Gaussian for tone feature extraction are:

- Due to the directional property of the filter, the filter can smoothen the normalised ΔT_0.
- The decision is made based on information from several frames instead of one frame, therefore it is robust to gross errors of an individual frame in the group.
- The normalised ΔT_0 is calculated according to the position and the orientation of the filter that produce the maximum output. Therefore, the proposed method does not require a pitch tracking algorithm and is robust to doubled/halved T_0.
- The oriented filter is steerable, i.e. the output of the filter can be expressed as a weighted sum of outputs of the designed basis filters. Therefore, the computation efficiency is high.

B. Sharpened weighted-autocorrelation function: Autocorrelation function is robust to noise, but the formant interference is the main problem for pitch determination. The weighted autocorrelation analysis [6] can suppress the formant structure and provide better performance on pitch determination than the conventional autocorrelation analysis, thus it is a good idea to apply the 2D oriented filter on the weighted autocorrelation function to produce a peak in the pitch orientation at T_0. The weighted autocorrelation function is defined as [6]

$$r(\tau, t) = \phi(\tau, t)/[\psi(\tau, t) + 1] \tag{4}$$

where τ is time lag, t is the frame index, $\phi(\tau, t)$ is the autocorrelation and $\psi(\tau, t)$ is the average magnitude difference function. $\phi(\tau, t)$ and $\psi(\tau, t)$ are given by

$$\phi(\tau, t) = \frac{1}{N}\sum_{n=0}^{N-1} s(n + t * T_s)s(n + t * T_s + \tau) \tag{5}$$

$$\psi(\tau, t) = \frac{1}{N}\sum_{n=0}^{N-1} |s(n + t * T_s) - s(n + t * T_s + \tau)| \tag{6}$$

where $s(n)$ is the speech signal and T_s is the frame step.

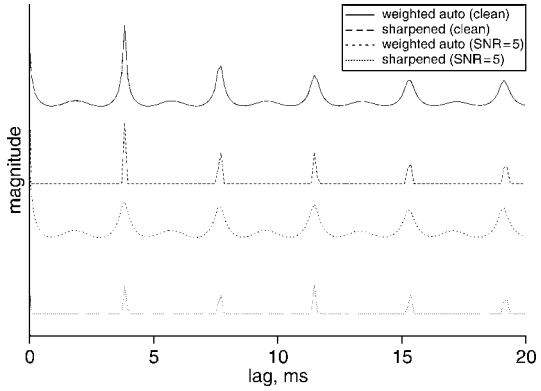

Fig. 1 *Original and sharpened weighted autocorrelation functions*

However when the signal-to-noise ratio (SNR) becomes lower, pitch harmonic peaks become less sharp, as shown in Fig. 1. On the other hand, the 2D oriented filter requires sharp peaks in order to obtain an accurate estimation of dynamic tone features. To solve this problem, a nonlinear filtering method is used to sharpen peaks and further reduce noise:

$$\tau_{sd}(\tau, t) = \max\left\{\sum_k g_2(k)r(\tau - k, t) - \left|\sum_k g_1(k)r(\tau - k, t)\right|, 0\right\} \tag{7}$$

where $r_{sd}(\tau, t)$ is the sharpened weighted-autocorrelation function, and $g_2(\tau)$ and $g_1(\tau)$ are given by

$$\begin{aligned} g_2(\tau) &= (1 - 2\tau^2/\sigma_\tau^2)e^{-\tau^2/\sigma_\tau^2} \\ g_1(\tau) &= (\tau^2/\sigma_\tau^2)e^{-\tau^2/\sigma_\tau^2} \end{aligned}$$

and σ_τ is a constant. Fig. 1 clearly shows that the sharpened weighted-autocorrelation function is ready for dynamic tone feature extraction as its peaks are very sharp even for noisy speech.

C. Dynamic tone feature extraction: As speech is a time-variant signal, we can only use few neighbouring frames for tone feature extraction. At the same time, a relatively large range of time lag should be used to cover possible values of ΔT_0. Therefore different scales are defined as:

$$x = \tau/\sigma_x \qquad y = t/\sigma_y \tag{8}$$

where σ_x and σ_y are scaling constants of τ and t, respectively. Now, we can apply the oriented filter on $r_{sd}(\tau, t)$. Let $O_2^\theta(x, y)$ represent the output of $G_2^\theta(x, y)$. As a output of $G_2^\theta(x, y)$ at a pitch harmonic peak should be maximised when the orientation of the filter matches the harmonic orientation, we can estimate the normalised ΔT_0 by solving $(\partial O_2^\theta(x, y)/\partial\theta) = 0$:

$$\widetilde{\Delta T_0} = -\frac{1}{T_p}\tan(\theta_d)\frac{\sigma_x}{\sigma_y}/T_s \tag{9}$$

where the maximum output occurs at T_p and orientation θ_d.

Some experimental results: To evaluate the performance of the proposed method, *Keele Pitch Database* [7] was used. In the system, speech was lowpass filtered with cutoff frequency of 700 Hz and then down-sampled to 8000 samples per second. The parameters used in experiments are $\sigma_\tau = 10$, $\sigma_x = 7$ and $\sigma_y = 2.5$ and $T_s = 10$ ms. Fig. 2 shows an example of the dynamic tone information extracted from the true pitch and from the clean speech by the proposed method. Owing to irregular excitation, the true $\widetilde{\Delta T_0}$ is not smooth. In contrast, the $\widetilde{\Delta T_0}$ extracted by the proposed method is much smoother and this is desirable for tone language speech recognition.

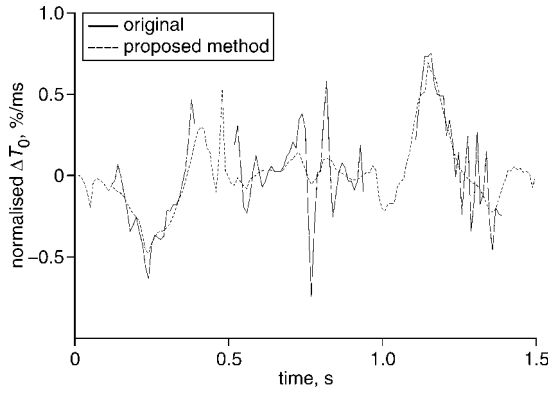

Fig. 2 *Smoothness of extracted normalised-ΔT_0*
——— $\widetilde{\Delta T_0}$ derived from *Keele Pitch Database*
---------- $\widetilde{\Delta T_0}$ extracted by proposed method

The proposed method is also robust to noise. Fig. 3 shows examples of $\widetilde{\Delta T_0}$ extracted from speech corrupted by additive white Gaussian noise. Clearly if $SNR \geq 5$, the $\widetilde{\Delta T_0}$ extracted from noisy speech is almost identical to that from clean speech in voiced frames. Even when $SNR = 0$, the proposed method can give good estimation in most frames, and significant errors only occur at very weak speech frames.

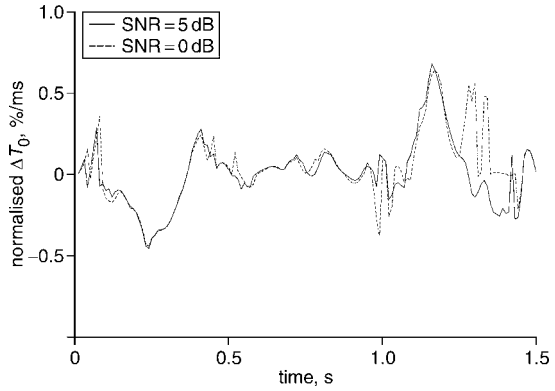

Fig. 3 *Robustness of proposed method*
——— $\widetilde{\Delta T_0}$ extracted from noisy speech at $SNR = 5$ dB
--------- $\widetilde{\Delta T_0}$ extracted from noisy speech at $SNR = 0$ dB

Conclusions: A novel method for dynamic tone feature extraction is presented. As $\widetilde{\Delta T_0}$ is obtained without knowing pitch, it can be used to improve pitch tracking algorithms for conventional PDAs. The proposed method uses an oriented filter which utilises information from neighbouring frames of speech, therefore it is robust to noise. Because pitch orientation is obtained by a group decision, the proposed method is robust to failures of individual frames and the extracted ΔT_0 is smooth. As the proposed method is designed to find the directions of pitch harmonic changes, it is inherently insensitive to the problem of halved/doubled pitch. The quality of tone features produced by the proposed method is remarkably good.

12 October 2004
Electronics Letters online no: 20057382
doi: 10.1049/el:20057382

J. Cai (*Queensland University, School of SEDC, 2 George Street, GPO Box 2434, Brisbane, Queensland 4001, Australia*)

References

1 Chen, S.-H., and Wang, Y.-R.: 'Tone recognition of continuous Mandarin speech based on networks', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (2), pp. 146–150
2 Demeechai, T., and Mäkeläinen, K.: 'New method for mitigation of harmonic pitch errors in speech recognition of tone languages'. Proc. IEEE Nordic Signal Processing Symp., 2000, pp. 303–306
3 Tabrikian, J., Dubnov, S., and Dickalov, Y.: 'Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model', *IEEE Trans. Speech Audio Process.*, 2004, **12**, (1), pp. 76–87
4 Cheveigné, A., and Kawahara, H.: 'Comparative evaluation of F0 estimation algorithms'. Proc. of Eurospeech, 2001, pp. 2451–2454
5 Freeman, W.T., and Adelson, E.H.: 'The design and use of steerable filters', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991, **13**, pp. 891–906
6 Shimamura, T., and Kobayashi, H.: 'Weighted autocorrelation for pitch extraction of noisy speech', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (7), pp. 727–730
7 Plante, F., Ainsworth, W.A., and Meyer, G.: 'A pitch extraction reference database'. Proc. of Eurospeech, Madrid, Spain, 1995, pp. 837–840