Pitch Estimation Using Mutual Information

Majid Mirbagheri[1], *Yanbo Xu*[1], *Shihab Shamma*[1,2]

[1]Institute for System Research, University of Maryland College Park, MD, USA
[2]Department of Electrical and Computer Engineering, University of Maryland College Park, MD, USA

mbagheri@umd.edu, yanbohsu@umd.edu, sas@umd.edu

Abstract

A spectrotemporal method based on Mutual Information (MI) is proposed for pitch estimation of voiced speech signals. We use MI as the similarity measure between voiced speech segments and their delayed version. Instead of measuring linear dependencies, MI measures statistical dependency, which suits the dynamic characteristic of speech signals. Besides, higher-order statistics are directly encoded in the MI while they are not usually taken into account in traditional correlation-based measures. Through experiments on both synthetic signals and a real speech dataset, this new measure is proven to be effective for pitch estimation.

Index Terms: pitch estimation, mutual information, higher-order statistics, periodicity, spectrotemporal

1. Introduction

Pitch is a significant attribute of voiced speech signals, and its accurate estimation plays a role of great importance for various applications like speech coding, enhancement, recognition, and so forth. The passed decades have witnessed the steady progress in pitch estimation algorithms. Basically, these algorithms can be classified into 3 categories: temporal, spectral, and spectrotemporal approaches [1].

As a classic model for temporal and spectrotemporal approaches, the autocorrelation function [2] measures the correlation of the windowed N-sample speech segment starting at t and its lagged version by τ, based on the assumption that voiced speech signal is quasi-periodic.

$$R(\tau, t) = \frac{1}{N} \sum_{n=t}^{t+N-1} x_n x_{n+\tau} \qquad (1)$$

Several methods [3] [4] have been proposed to modify the correlation in (1) in order to improve the accuracy. However, these methods only make use of information up to the second-order statistics of the analyzed signal. To further exploit the higher-order statistics, Wu *et al.* [5] applied correntropy to analyze the temporal structure of speech signal, and explored the superiority of correntropy over traditional autocorrelation based methods. Given a one-dimensional signal with N available sample pairs $(x_n, x_{n+\tau})$ the correntropy-based autocorrelation function is estimated by:

$$R_V(\tau) = \frac{1}{N} \sum_{n=1}^{N} G(x_n - x_{n+\tau}) \qquad (2)$$

where $G(x)$ is Gaussian kernel $\exp(-\frac{x^2}{2\sigma^2})$. The Taylor Series expansion of the correntropy function indicates that only the even-order moments of the error $x_n - x_{n+\tau}$ are included meaning that as a function of the squared difference error it reflects the the similarity of samples spaced at different lags. Both correntropy and correlation-based measures capture the linear dependencies between these samples. The issue about linear dependency is that it cannot be relied on in some realistic scenarios where the periodicity is implicitly manifested in a transformed version of its signal such as envelope as opposed to explicitly existing in the signal itself. To address this issue, we propose using statistical dependence of samples and their lagged versions by measuring mutual information between them. The paper is organized as follows. In section 2, we discuss the use of MI to estimate the pitch of speech signals, and in section 3 evaluate the performance of the proposed method. Finally we discuss the implications of this work in section 4.

2. Method

We first introduce Mutual Information and its estimation. We also discuss how we benefit from MI to capture the periodicity of one-dimensional signals, and describe the the pitch detection algorithm for speech signals based on MI.

2.1. Mutual Information

Mutual Information (MI) is a measure of statistical dependence between random variables. MI has previously been used to measure similarity in the context of clustering and feature selection in [6, 7]. For two continuous random variables X and Y with joint and marginal densities $f_{X,Y}$, f_X and f_Y, Shannon's MI is defined as:

$$I(X,Y) = \iint f_{X,Y}(x,y) \ln \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, \mathrm{d}x \, \mathrm{d}y \qquad (3)$$

It can be shown that $I(X, Y) \geq 0$ and $I(X, Y) = 0$ *if and only if* X and Y are independent. MI is unique in its close ties to Shannon entropy and the theoretical advantages derived from this. With the entropy as a measure of the uncertainty associated with random variable X and defined as:

$$H(X) = -\int f_X(x) \ln f_X(x)\, \mathrm{d}x \tag{4}$$

MI measures how much knowing one of the variables reduces the uncertainty about the other or mathematically:

$$I(X, Y) = H(X) - H(X|Y) \tag{5}$$

In applications, the joint densities are usually unknown and one has the data available in form of N sample points (x_i, y_i), $i = 1, \ldots, N$ which are assumed to be i.i.d. realizations of the underlying joint density $f_{X,Y}$. Among numerous existing algorithms to estimate $I(X, Y)$, we chose a k-nearest neighbor (KNN) estimator introduced in [8] previously shown to be fairly accurate and data-efficient [9].

This method gives a nonparametric estimation of mutual information between two random X and Y based on i.i.d. samples $z_1, \ldots, z_N$, $z_i = (x_i, y_i)$. Using max norm for the space Z defined as:

$$\rho(z_i, z_j) = \max\{D_x(x_i, x_j), D_y(y_i, y_j)\} \tag{6}$$

in which D_x and D_y can be any two arbitrary distance functions defined on X and Y subspaces, the KNN estimator of mutual information would be derived as followed:

$$I_{k,N}(X, Y) = \psi(k) - \frac{1}{k} - \frac{1}{N} \sum_{i=1}^{N} (\psi(l_{i,k}^x) + \psi(l_{i,k}^y)) + \psi(N) \tag{7}$$

with ψ being the famous digamma function defined as:

$$\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \frac{d}{dz} \ln \Gamma(z) = \int_0^\infty (\frac{e^{-t}}{t} - \frac{e^{-zt}}{1 - e^{-t}}) dt \tag{8}$$

and $\rho_{i,k}^z$ being the distance between z_i and its k-th nearest neighbor among $N - 1$ remaining samples according to (6), and $\rho_{i,k}^x$ and $\rho_{i,k}^y$ the distances between the same points projected into the X and Y subspaces ($\rho_{i,k}^z = \max\{\rho_{i,k}^x, \rho_{i,k}^y\}$), $l_{i,k}^x$ and $l_{i,k}^y$ are defined as the number of samples with $D_x(x_i, x_j) \leq \rho_{i,k}^x$ and $D_y(y_i, y_j) \leq \rho_{i,k}^y$. For our application, D_x and D_y were chosen to be the max norm and $\sqrt{k/N} = 0.4$ as advised in the method implementation.

2.2. Periodicity Estimation

In order to detect structural periodicities in one-dimensional signals, we use estimates of MI between their samples and delayed version ones. For that we define the autodependence function for one-dimensional signal x, $R_I(\tau)$, by:

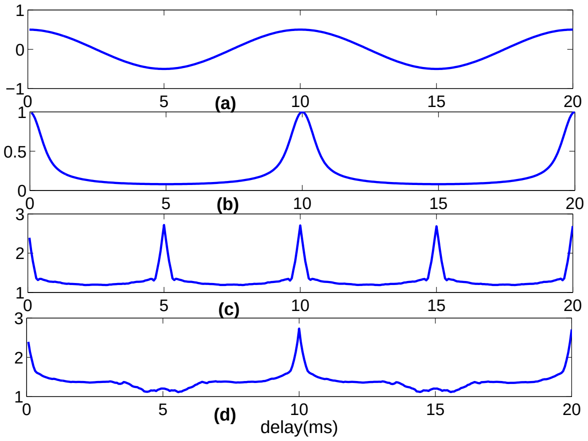

Figure 1: *(a) $R(\tau)$, (b) $R_V(\tau)$, (c) $R_I(\tau)$ calculated for a 100Hz sinusoid of duration 20ms, (d) $R_I(\tau)$ calculated for the half-rectified, low-passed version of the same signal*

$$R_I(\tau) = I_{k,N}(X_0, X_\tau) \tag{9}$$

with X_τ denoting the random variable realized by N samples of signal x delayed by the lag τ. Now we look at how this function behaves for different synthetic signals and compares it to autocorrelation functions based on correlation and correntropy, $R(\tau)$ and $R_V(\tau)$. Throughout the section, we use the notation $R(\tau)$ without the index t for that in all cases only one single window was applied to the sample segments.

The first signal we illustrate is a simple 100Hz sinusoid of duration 20ms sampled at 16kHz. The three top plots in figure 1 show the values of $R(\tau)$, $R_V(\tau)$ and $R_I(\tau)$ computed for this signal. Comparing the peaks at $\tau = 10ms$, it can be seen that $R_I(\tau)$ gives rise to a much sharper peak. The sharpness of the peak is advantageous in applications dealing with multiple pitches. The undesired peak in $R_I(\tau)$ at $\tau = 5ms$, half of the actual period T is a direct result of the relation $x(t) = -x(t + T)$ for the pure sinusoid signals. This unwanted peak at $\tau = \frac{T}{2}$ can be easily removed by half-rectifying and low-pass filtering of the signals before computing the R_I function as shown in the bottom plot in the same figure.

In the next example, we explore the case when the 100Hz sinusoid signal was modulated at 361Hz. This specific frequency was deliberately chosen a coprime to the frequency of the original sinusoid so that the two tones did not beat together. Figure 2 shows the $R(\tau)$, $R_I(\tau)$ and $R_V(\tau)$ values computed for delays in the range 0-20 ms. As shown in the plots, the desirable peak at the delay $\tau = 10ms$ corresponding to the envelope periodicity is only observable in R_I. This can be related to the fact

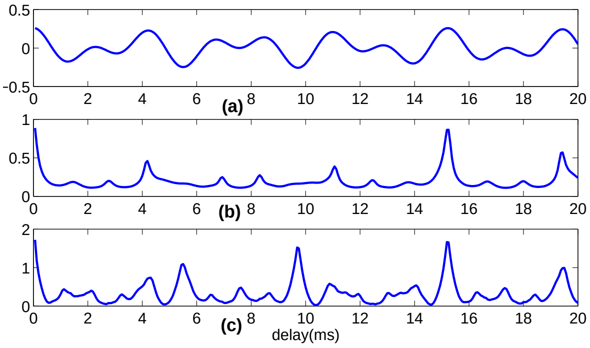

Figure 2: *(a) $R(\tau)$, (b) $R_V(\tau)$, (c) $R_I(\tau)$ calculated for a 100Hz sinusoid modulated at 361Hz*

that the coprime modulator has disrupted the linear dependence between samples spaced at the period of the envelope while they still maintain a relatively high statistical dependence as reflected in R_I. This phenomenon is of particular importance for spectrotemporal algorithms when dealing with real-world dynamical signals such as speech for which: 1) periodic segments are amplitude-modulated by other slow-varying signals 2) periodicities in subband channels are manifested in the envelope of the filters outputs.

2.3. MI-based Pitch Estimation

In this application, we include a feature analysis stage using a gammatone filter bank which mimics the function of human cochlea, and is widely used as a peripheral processing module in computational auditory scene analysis (CASA) models. The input signal is separated into frequency bands while the temporal structure of the original signal is preserved. The filterbank consists of constant-Q bandpass filters [1], with high resolution in the the lower frequency channels so as to separate the first several harmonics which serve as important cues for pitch estimation. Thus this method is cast into a spectrotemporal framework. Once analyzed, the output of each filter is halfway rectified, and then low-passed mimicking the envelope tracking function of human inner hair cells. Assuming that channels dominated by harmonics resemble the sinusoidal signal in figure 1, the halfwave rectification then helps in suppressing unwanted peaks in our pitch estimation method.

After the peripheral processing, segments of subband signals are extracted and the autodependence functions are computed for each and denoted by $R_I^i(\tau)$, for $1 \leq i \leq 64$. Instead of pooling these functions across subbands, we use weighted sums with different weighting rules specifically designed for different time lags. Denoting the sampling frequency of the original speech waveform as f_s, given a time lag τ, for the hypothesized fundamental frequency, $f_\tau = f_s/\tau$ different harmonics of the fundamental frequency at multiples of f_τ are considered. As the bandwidth of channels might be relatively broad, each channel can span several neighbor harmonics. For the sake of simplicity, assuming that the output of each channel is dominated by the nearest harmonic to its center frequency in log-scale, the weight of the i-th channel, w_i^τ is determined in the following way:

$$w_i^\tau = \cos\left(\frac{\pi(f_{ci} - f_{n_i})}{f_\tau}\right) \tag{10}$$

with f_{ci} denoting the center frequency of the i-th channel and f_{n_i} the nearest harmonic to f_{ci}. The logic behind this is that the closer a harmonic to the center frequency of the channel, more likely this channel contains relevant information about the pitch candidate. Finally we have the final formula for the aggregated autodependence as a function of the time lag τ as:

$$\mathcal{R}_I(\tau) = \sum_{i=1}^{64} R_I^i(\tau) w_i^\tau \tag{11}$$

3. Experimental Results

To demonstrate the accuracy of the proposed MI-based method in resolving multiple pitches we applied it on a mixture of two synthetic vowels /a/ and /i/ with close pitch values of 100Hz and 105Hz. Depicted in figure 3 it can be seen that the sharper peaks at the corresponding delays in autodependence function (specially compared to the correlation-based one) makes it easier to reliably detect both pitch values. For better clarity, the bottom two plots are normalized so that the minimum and the maximum become 0 and 1. We also assessed the effectiveness of MI as the similarity measure for pitch estimation on Bagshaw's FDA dataset [10]. This dataset is composed of speech signals from one male and one female

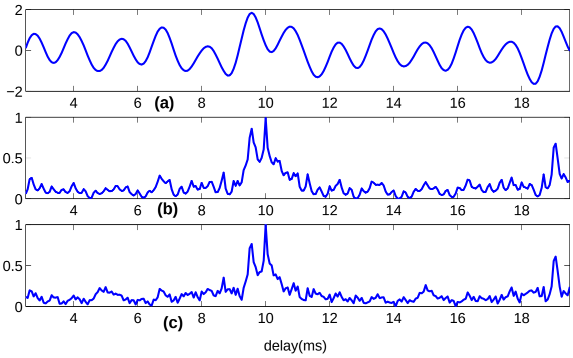

Figure 3: *(a) $R(\tau)$, (b) $R_V(\tau)$, (c) $R_I(\tau)$ calculated for a mixture of two synthetic vowels /a/ and /i/ with pitch values of 100Hz and 105Hz*

SNR(dB)	clean	10	0
White Noise Added			
MI	5.72	6.18	12.18
CORRE	11.93	13.54	22.15
AUTO	10.72	10.96	13.48
Babble Noise Added			
MI		13.62	37.66
CORRE		22.89	42.55
AUTO		14.07	30.48

Table 1: Comparison of GPE for MI, CORRE, and AUTO pitch estimation methods

speaker, and the spontaneous laryngograph signals. The same 50 sentences are read by each speaker and recorded. Although files containing reference of pitch contours are also provided, we regenerate ground truth by running an autocorrelation method on the laryngograph signal with manual correction to align it to the right time resolution. White and babble noise from NOISEX92 dataset were added at two different SNR levels 10, and 0dB to simulate noisy conditions.

We compared our method (MI) with correntropy (CORRE) [5] and classical autocorrelation (AUTO) as they share similar underlying mechanisms for pitch estimation. Pitch values were determined from speech segments of length 20ms. The pitch range considered was from 50Hz to 400Hz, corresponding to a lag range from 40 to 320 samples. Thus in total, 40ms (640 samples) was used for the analysis at each time step. The same channel weighting was applied for all three measures. As pointed out in [11], the accuracy of determination of boundaries of voiced speech sessions can affect the performance of pitch determination algorithms. Since the focus of this paper is on the similarity measure of voiced speech segment, only the segments within voiced sessions were analyzed according to the obtained ground truth of pitch contours. It should be also noted that different methods might apply different post-processing procedures to refine the results, which usually require parameter adjustment by trial and error. Hence, to compare these 3 measures, no post-processing was conducted, and for each time step only the maximum peak within similarity values computed at different lags was chosen. We adopt Gross Pitch Error (GPE) [12] to determine the correctness. If the determined pitch is more than 20% off the reference pitch value, we consider an error. The error rates for the 3 methods under different conditions are summarized in table 1. The error rates are higher than previously reported ones because of the lack of the post-processing stage. From table 1, MI-based method has the lowest error rate for clean and white noisy signals. The performance for all these methods degrade as the SNR drops, especially for babble noise. For this noise type, our method outperforms the other two at SNR of 10dB, but becomes slightly weaker than AUTO when SNR is lower. This can be explained in the shadow of the fact that in presence of a dominant highly non-stationary noise higher order statistics of the signals becomes less representative of their nature and hence less reliable.

4. Conclusion

In this paper, we proposed to use Mutual Information as a new similarity measure of voiced speech segments with their delayed versions for pitch estimation. This measure is distinguished by its use of statistical dependency measures instead of the limited linear dependency. Through experiments with both synthetic data and a real speech dataset, this new measure was shown to outperform the traditional measures that rely on information from second order statistics. We believe with post-processing refinement, the performance of our method could be further improved.

5. References

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principle, Algorithms, and Applications*. Wiley, 1994.

[2] P. Boersma, “Praat, a system for doing phonetics by computer.” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[3] J. C. Brown and M. S. Puckette, “Calculation of a “Narrowed” Autocorrelation Function,” *Journal of Acoustical Society of America*, vol. 85, no. 4, pp. 1595–1601, 1989.

[4] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] J. W. Xu and J. C. Principe, “A Pitch Detector Based on a Generalized Correlation Function,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1420–1432, 2008.

[6] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, “Normalized mutual information feature selection,” *Neural Networks, IEEE Transactions on*, vol. 20, no. 2, pp. 189 –201, feb. 2009.

[7] A. Kraskov, H. Stögbauer, R. Andrzejak, and P. Grassberger, “Hierarchical clustering using mutual information,” *EPL (Europhysics Letters)*, vol. 70, p. 278, 2005.

[8] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.

[9] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson, V. Protopopescu, and G. Ostrouchov, “Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data,” *Phys. Rev. E*, vol. 76, p. 026209, 2007.

[10] B. Hiller, S. M. Hiller, and M. A. Jack, “Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching,” in *in Proceedings of the 3rd European Conference on Speech Communication and Technology*, 1993, pp. 1003–1006.

[11] W. Chu and A. Alwan, “Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend,” in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3969–3972.

[12] L. Rabiner, M. Cheng, A. Rosenberg, and C. 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.