FUNDAMENTAL FREQUENCY ESTIMATION BASED ON PITCH-SCALED HARMONIC FILTERING

Sergio Roa, Maren Bennewitz, and Sven Behnke

University of Freiburg
Department of Computer Science
{roa, maren, behnke}@informatik.uni-freiburg.de

ABSTRACT

In this paper, we present an algorithm for robustly estimating the fundamental frequency in speech signals. Our approach is based on pitch-scaled harmonic filtering (PSHF). Following PSHF, we perform a filtering in the frequency domain using the short-time Fourier transform in order to separate the harmonic and non-harmonic parts of the processed signal. We enhance the standard PSHF approach by using a range of window lengths and a cost function that is applied to each window size. This cost function takes into account the energy at the harmonic and non-harmonic frequency coefficients to estimate harmonic energy for a frame. By using energy peaks and applying a cost function that considers the change in pitch in subsequent frames, we then determine the final pitch contour. We evaluated our approach on the Keele database. As the experimental results demonstrate, our methods performs robustly for noisy speech and has a good performance for clean speech in comparison with state-of-the-art algorithms.

Index Terms— Speech processing, pitch estimation.

1. INTRODUCTION

The human brain performs a kind of feature extraction and pattern recognition from the original speech signals at the level of the harmonic complexes. Thus, a human being is able to direct its attention to any desired speech source, despite no ideal acoustic conditions. The fundamental frequency (F_0) or pitch of the human voice plays a central role in both the production and perception of speech [1]. In clean and corrupted speech, pitch is generally perceived with high accuracy at the fundamental frequency characterizing the vibration of the speaker's vocal chords [2].

The problem of pitch estimation has been addressed for a long time. Many different approaches have been proposed and there is large availability of literature on machine algorithms for pitch tracking [3]. Recently, techniques like statistical learning for feature extraction [4, 5], time domain probabilistic approaches for waveform analysis [6], and optimization techniques [7, 8] have been presented. However, most of these methods lack robustness, especially for corrupted speech.

In this paper, we present a technique to robustly estimate pitch in speech signals. Our approach is based on pitch-scaled harmonic filtering (PSHF). PSHF has been designed to separate the periodic and aperiodic components of speech signals [9]. PSHF uses the short-time Fourier transform and different window sizes to evaluate the periodicity of the signal, expecting to match a given window length to a multiple of the pitch period. In contrast to the original PSHF that uses window sizes scaled to a given multiple of pitch periods and applies an optimization technique to find the optimum window size for each section of speech, we apply a range of window lengths according to a range of multiples of pitch periods at equidistant points in time and then we use a cost function and a dynamic programming method to track a pitch contour. This resulted in an accurate pitch estimation and a good compromise between time and frequency resolution. To estimate the presence of a range of F_0-frequencies at each frame, we use a cost function called harmonic template function (HTF). The HTF takes into account the energy at the harmonic and non-harmonic frequency coefficients to estimate harmonic energy for a frame. As a result, we get a vector corresponding to the range of window sizes that contains energy peaks which can be used for pitch tracking. To find the final pitch contour, we integrate the harmonic energies found for each time step of processing by applying a cost function that considers the change in pitch in subsequent HTF vectors.

This article is organized as follows. In the following Section, we introduce PSHF. In Section 3, we present the Harmonic Template Function and in Section 4 we explain how to find pitch contours. Finally, in Section 5 we present experiments on the Keele database showing the robustness of our technique for noisy speech and the good performance for clean speech in comparison with state-of-the-art algorithms.

2. PITCH-SCALED HARMONIC FILTERING

Using spectrograms is common to compute features that represent the spectral envelope [3]. This approach has two draw-

This work has been supported by the German Science Foundation (Deutsche Forschungsgemeinschaft, DFG), grant BE 2556/2-1.

backs in the context of pitch estimation. First, the smooth windows have a frequency response themselves, leading to smearing of the harmonic energy over multiple neighboring frequency coefficients (bins). This phenomenon is called spectral leakage. Second, since the length of the analysis window is predetermined, the multiples of the fundamental usually fall between frequency bins. Hence, the estimation of peak heights is made difficult. One possibility is to use a rectangular window whose length is exactly one pitch period, which is called pitch-synchronous analysis. A smooth window can also be used, but it must cover exactly two pitch periods. The difficulty in computing pitch-synchronous analysis is that we need to know the local pitch period [3]. To avoid the problems induced by analysis windows, Jackson and Shadle [9] proposed a method called pitch-scaled harmonic filter. The algorithm presented here is based on this procedure.

The PSHF uses the spectral properties of an analysis frame scaled to the pitch period in order to distinguish harmonic and non-harmonic parts of the spectrum and hence to arrive to a pitch period estimate. The term pitch-scaled refers to an analysis frame that contains a small integer multiple of pitch periods. The advantage of this property is that the harmonics of the fundamental frequency F_0 will be aligned with certain bins of the short-time Fourier transform (assuming we know F_0). For example, if the analysis frame contains b pitch periods, then the frequency of the ibth Fourier coefficient will correspond to iF_0. More specifically, a window function w centered at time m of length N is applied to the speech signal $x(n)$ to form $x_w(n) = w(n)x\left(n+m-\frac{N}{2}\right)$. The discrete Fourier transform (DFT) is used to compute the spectrum $X_w(k,m) = \sum_{n=0}^{N-1} x_w(n)e^{-j\frac{2\pi}{N}kn}$ by using a value of $N = bP$ where b is the number of pitch periods of length P (in samples) [9]. Hence, the periodic part of x_w is concentrated into the set of harmonic bins $\mathcal{B}$, where $\mathcal{B} = \{b, 2b, 3b, \ldots, (N-1)b\}$. If the length of the analysis window is matched to a multiple of the pitch cycle, the harmonic energy is centered at these frequency bins. Since the signal is now aligned, one can apply a rectangular window, instead of a smooth window. The intermediate bins will therefore contain energy that is either non-harmonic or has a different fundamental frequency. Hence, we can assume that these bins correspond to the local noise level. In [9], Hanning windows and four-pitch period windows were used. For each section of speech, an optimization procedure was implemented which estimates the calculated and measured smearing by using specific window sizes and then minimizes the difference between both quantities using a given cost function. This way, the optimum window size is found and can be used to estimate harmonic and non-harmonic spectra. In contrast, in this work a range of window lengths and pitch periods were used to estimate F_0.

As an example for PSHF, Figure 1 shows the spectrum for a vowel /a/ spoken by a male with $F_0 = 110$ Hz (pitch period 9.1 ms), and corresponding harmonic and non-harmonic spectra, computed using a four-pitch period window.

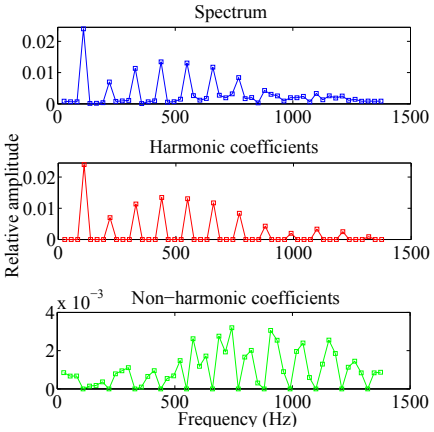

Fig. 1. Pitch-scaled harmonic filtering.

3. THE HARMONIC TEMPLATE FUNCTION

We use different window lengths applied at equidistant points in time as a local indicator of the existence of pitch and in such a case, as an estimator of the fundamental. We can estimate the harmonic energy by maximizing the difference between the energy at harmonic bins and the level of noise at the remaining bins, according to a cost function. We call this cost function a harmonic template function (HTF) and use it for pitch estimation. If the energy concentrated at harmonic bins is high enough relative to its neighboring bins, the magnitudes of DFT peaks provide a good estimate of the current pitch. For a DFT of length N, we are interested in using the first non-redundant $N_h \approx \frac{N+1}{2}$ bins. The total number of expected harmonic bins is $H \approx \frac{N_h-1}{b}$. Let us define X_h^{-d} and X_h^{+d} as the energy of the surrounding bins for a given window at each harmonic $h \in \{1, 2, \ldots, H\}$ and at a bin-distance $d = 1, \ldots, b-1$. For a four-pitch period window the estimated noise is concentrated at three surrounding lower-frequency bins and three higher-frequency bins.

The function that maps the window length to its corresponding estimated fundamental frequency is nonlinear. As a consequence, higher errors in the estimated F_0 for high frequencies are expected when the pitch tracking algorithm is applied. In order to reduce the repercussion of scaling and to attain an improved F_0 estimation, we designed a cost function HTF that combines different values b of pitch periods. The most convenient design of the HTF is an open question. We use the binomial coefficients to automatically generate appropriate HTFs. Thus, the goal is to apply a positive weight for the estimated harmonic bin and a negative weight for the estimated non-harmonic bins, which should be proportional to the positive one. Hence, for $b = 2, \ldots, 10$, the cost function for the surrounding bins of a given harmonic h is defined in general as

$$S_h^b(N,m) = \sum_{d=1}^{b-1} \frac{1}{K} \binom{2(b-1)}{d-1} \left(|X_h^d| + |X_h^{-d}|\right) \qquad (1)$$

Fig. 2. (a) Harmonic template function for pitch estimation using 2 to 10-pitch period windows of length 400 to 800 for the utterance "The north" (male). (b) Final HTF after applying feature extraction operations. The circles show the path found by the pitch tracking algorithm.

and the HTF as

$$J^b(N,m) = \sum_{h=1}^{H} U_h^b(N,m) = \sum_{h=1}^{H} |X_w(bh+1)| - S_h^b, \quad (2)$$

where K is a normalization constant for generating relative values for all HTFs. Each HTF maximizes the accumulated energy found at the harmonic bins and serves as an estimate of the pitch. If $J^b(N,m) < 0$, we set $J^b(N,m) = 0$. This occurs when the window length does not match the correct pitch period. Note that the frequency ranges mostly overlap for two or more different values of b. Therefore, a suitable interpolation of the values obtained from each HTF J^b is required. To avoid inconvenient nonlinearities in the interpolation, we use a logarithmic scale for the range of F_0 frequencies, starting from 50 Hz to 500 Hz. 1536 interpolation indices were generated. Thus, the values returned by the HTFs were interpolated and normalized, by using a linear interpolation. Additionally, a mean filter was applied to reduce improper discontinuities at the borders of the overlapping areas. We analyze speech segments sampled at $F_s = 20$ kHz. Each signal is processed every 5 ms. For $b = 2, \ldots, 10$, window lengths ranging from 400 to 800 samples were used, allowing us to span the range of frequencies from 50 Hz to 500 Hz. The result can be observed in Figure 2(a).

As expected, for certain frequencies being factors or multiples of the fundamental, energy peaks are also found. These frequencies correspond to pitch intervals, like octaves $(2:1)$ or fifths $(3:2)$. These peaks have a significant impact on the performance of a pitch tracking algorithm. To deal with these issues, we apply two different operations. The first problem appears in the case in which window lengths correspond to half of the fundamental, where peaks of energy are immediately followed by valleys. Given a function U_h^b, we apply the following linear operation to eliminate these peaks, by using the adjacent values U_{h+1}^b and U_{h-1}^b:

$$U_h^b(N,m) = \min\left\{U_{h+1}^b + U_{h-1}^b, U_h^b\right\} \quad (3)$$

for $h = 2, \ldots, H-1$. After interpolating the results of all HTFs, we apply the second operation. Assuming that the highest peaks are located on the estimated fundamental, we subtract half of the energy contained at F_0 from the energy at its multiples. The final HTF, defined as $J(i,m)$, where $i = 1, \ldots, 1536$ and m is the current time frame, results from such operations. Figure 2(b) shows the final HTF resulting from the feature extraction process applied to the male speech signal of Figure 2(a). The energy was erased up to a given F_0 at multiples $2F_0$ and $3F_0$.

4. PITCH TRACKING

We use the fact that the pitch does not change abruptly to design a pitch tracking algorithm that uses dynamic programming techniques. The algorithm finds a compromise between maximizing the energy at the frame m and optimizing the shift from a contiguous frame. Let us define $\mathbf{J} = \{\mathbf{J}(1), \ldots, \mathbf{J}(M)\}$ a sequence of HTF vectors which track the pitch for M consecutive frames. The maximum a posteriori (MAP) estimate of the pitch track is $\mathbf{J}_{MAP} = \max f(\mathbf{J})$. The function $f(\mathbf{J})$ constitutes a priori statistics for the pitch and can help disambiguate the pitch, by avoiding pitch doubling or halving given knowledge of the speaker's average pitch, and by avoiding rapid transitions given a model of how pitch changes over time. One possible approximation is given by assuming that the probability of the pitch period at frame m depends only on the pitch period for the previous frame [3]:

$$f(\mathbf{J}) = f(\mathbf{J}(M)|\mathbf{J}(M-1)) \cdots f(\mathbf{J}(2)|\mathbf{J}(1)). \quad (4)$$

This cost function f penalizes the current pitch estimation by using the distance to an accumulated maximum $\mu(j)$ along the sequence of HTF vectors, for a given index j (relative to some F_0). Moreover, the penalty considers the fact that the pitch does not change strongly, by defining a maximum shift constant δ. This way, we can follow the optimal path between $\mathbf{J}(m)$ and $\mathbf{J}(m+1)$ by identifying the optimal shift as we move forward. Finally, the pitch tracking algorithm has to decide if a frame is voiced or unvoiced. This is done by measuring the average HTF energy at a frame and its neighborhood and by measuring the average HTF pitch energy at the position given by the path and its neighboring path lo-

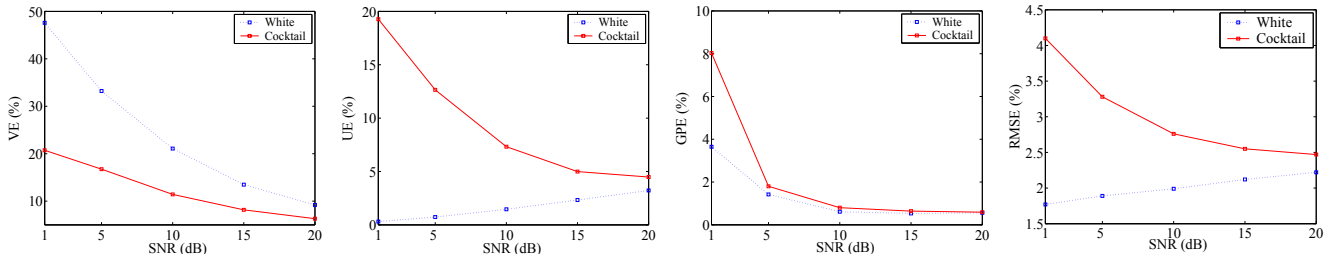

Fig. 3. Results obtained by using the Keele database signals mixed with cocktail party noise and white noise at different SNRs.

cations. If the ratio (pitch confidence) between the average pitch energy and the average energy exceeds a given threshold, the algorithm reports a voiced frame. In Figure 2(b) the estimated path is shown, where pitch reference values of unvoiced sections were set conventionally to 0, so as to use them as a reference to detect voiced frames.

5. EXPERIMENTAL RESULTS

We used the Keele pitch reference database [10] to evaluate the performance of the algorithm. It consists of speech signals of five male and five female English speakers, recorded with a sampling rate of 20 kHz and a resolution of 16 bit. For evaluation we used the pitch reference provided by Flego [11], who reanalyzed the original laryngograph signal to obtain improved pitch estimates every 1 ms. Common performance measures for comparing pitch estimation algorithms were used: The voiced error (VE) denotes the percentage of voiced time frames misclassified as unvoiced, the unvoiced error (UE) is defined as the inverse case, the gross pitch error (GPE) denotes the percentage of time frames at which the estimation and the reference pitch differ by more than 20%, and the root mean square error (RMSE) is computed as RMS difference in Hertz of the reference pitch and the estimation for all time frames that are not GPEs. The results for clean speech are shown in Tab. 1. The table also presents some comparative results obtained by Nonnegative Matrix Factorization (NMF) and RAPT algorithms [5]. To test the robustness of our approach, we added white noise and cocktail party noise (see Fig. 3) with different signal-to-noise ratios to the clean speech signals. For clean speech and moderated SNRs the results are reliable.

Table 1. Evaluation results obtained on the Keele pitch reference database for clean speech.

	VE (%)	UE (%)	GPE (%)	RMSE (Hz)
PSHF Based	4.51	5.06	0.61	2.46
NMF	7.7	4.6	0.9	4.3
RAPT	3.2	6.8	2.2	4.4

6. CONCLUSION

In this paper, we presented a PSHF-based approach for pitch determination. We enhanced PSHF by exhaustively using a range of window lengths and a cost function that allows to estimate the pitch for each frame. We track the pitch in subsequent frames using a cost function that penalizes larger changes in pitch. We evaluated our method for clean speech as well as for demanding acoustic conditions. The experimental results show the robustness of our method for noisy speech and the competitiveness with state-of-the-art algorithms for clean speech.

7. REFERENCES

[1] K. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1999.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, September 1994.

[3] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, 2001.

[4] F. Sha, J. A. Burgoyne, and L. K. Saul, "Multiband statistical learning for F0 estimation in speech," in *Proc. of ICASSP*, Montreal, Canada, 2004, pp. 661–664.

[5] F. Sha and L. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., pp. 1233–1240. MIT Press, Cambridge, MA, 2005.

[6] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segment-based probabilistic generative model of speech," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 2005, pp. 61–64.

[7] L. K. Saul, D. D. Lee, C. L. Isbell, and Y. LeCun, "Real time voice processing with audiovisual feedback: Toward autonomous agents with perfect pitch," in *Advances in Neural Information Processing Systems 15*, S. Thrun S. Becker and K. Obermayer, Eds., pp. 1181–1188. MIT Press, Cambridge, MA, 2003.

[8] L. Saul, F. Sha, and D. D. Lee, "Statistical signal processing with nonnegativity constraints," in *Proc. of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003, vol. 2, pp. 1001–1004.

[9] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, October 2001.

[10] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech '95*, J. M. Pardo et al., Eds. 1995, vol. 1, pp. 837–840, UP Madrid.

[11] F. Flego, *Fundamental Frequency Estimation Techniques for Multi-Microphone Speech Input*, Ph.D. thesis, University of Trento, Mar. 2006.