A Novel Two-Level Pitch Detection Approach for Speaker Tracking in Robot Control

Mahmoud R. Hejazi, Han Oh, Hong Kook Kim, and Yo Sung Ho

Department of Information and Communications,
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea
(E-mail: {m_hejazi,ohhan,hongkook,hoyo}@gist.ac.kr)

Abstract: Using natural speech commands for controlling a human-robot is an interesting topic in the field of robotics. In this paper, our main focus is on the verification of a speaker who gives a command to decide whether he/she is an authorized person for commanding. Among possible dynamic features of natural speech, pitch period is one of the most important ones for characterizing speech signals and it differs usually from person to person. However, current techniques of pitch detection are still not to a desired level of accuracy and robustness. When the signal is noisy or there are multiple pitch streams, the performance of most techniques degrades. In this paper, we propose a two-level approach for pitch detection which in compare with standard pitch detection algorithms, not only increases accuracy, but also makes the performance more robust to noise. In the first level of the proposed approach we discriminate voiced from unvoiced signals based on a neural classifier that utilizes cepstrum sequences of speech as an input feature set. Voiced signals are then further processed in the second level using a modified standard AMDF-based pitch detection algorithm to determine their pitch periods precisely. The experimental results show that the accuracy of the proposed system is better than those of conventional pitch detection algorithms for speech signals in clean and noisy environments.

Keywords: Pitch Detection, Cepstrum, AMDF, Neural Classifier, Speaker Verification, Human-Robot Controller

1. INTRODUCTION

The idea of using natural speech commands for controlling a human-robot is an interesting topic in the field of robotics [1], where the robot control system must determine if the person who gave the command is an authorized speaker, recognize what has been said, understand the command, and then relay the command to the appropriate robot system. A typical interface for such a purpose is illustrated in Fig. 1.

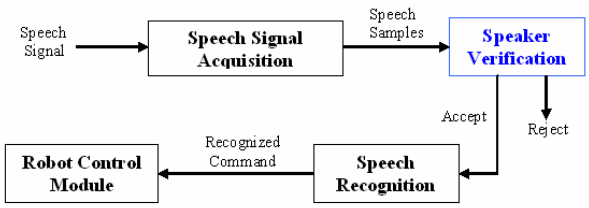

Fig. 1 A typical interface for controlling a robot by natural speech commands.

As shown in Fig. 1, the system consists of three different parts including speech acquisition, speaker verification, and speech recognition modules. Our main focus in this paper is to develop speaker verification module, where we use some characteristics and dynamic features of the human speech production system to verify whether or not the speaker is authorized for commanding to the robot. Among the possible dynamic features, we can name pitch period or fundamental frequency which is related to the voice excitation organs, and formants related to the human vocal tract [2] [3]. Usually, the pitch period characteristic is different from person to person and if determined precisely, it can be then used with other parameters such as speech spectrum envelope to verify a speaker.

Such an application as well as many other applications in the field of speech processing makes the pitch detection as a challenging topic for researchers [4]. However, current techniques of pitch detection are still not to a desired level of accuracy and robustness. When presented with a single clean pitched signal, most techniques do well, but when the signal is noisy, or when there are multiple pitch streams, the performance of the most developed techniques degrades.

In this paper, we present a two-level approach for pitch detection which in compare with the conventional pitch detection algorithms, not only increases the accuracy, but also makes the system more robust to noise. We will then show the application of our proposed approach for verification of speakers in a robot control system.

2. PITCH DETECTION ALGORITHMS

A pitch detector is an essential component in a variety of speech processing systems. Besides providing valuable insights into the nature of the excitation source for speech production, the pitch contour of an utterance is useful for recognizing speakers, for speech instruction to the hearing impaired, and is required in almost all speech synthesis systems.

However, accurate and reliable measurement of the pitch period of a speech signal from the acoustic waveform alone is often exceedingly difficult for several reasons. One reason is that the glottal excitation waveform is not a perfect train of periodic pulses. Although finding the period of a perfectly periodic waveform is straightforward, measuring the period of a speech waveform, which varies both in period and in the detailed structure of the waveform within a period, can be quite difficult.

A second difficulty in measuring pitch period is the interaction between the vocal tract and the glottal excitation. In some instances the formants of the vocal tract can alter significantly the structure of the glottal waveform so that the actual pitch period is difficult to detect. Another difficulty may arise in the practical situation where the speech signal is noisy. In such cases, the detailed structure of the waveform may be changed which leads to an incorrect measure of pitch period.

As a result of the numerous difficulties in pitch measurements, a wide variety of sophisticated pitch detection methods have been developed. Basically, a pitch detector is a device which makes a voiced-unvoiced decision, and, during periods of voiced speech, provides a measurement of the pitch

period, although in most algorithms, such decision-making is a part of the measurement process and not an individual one. Among different algorithms, the following categories are applied for pitch detection.

1) Autocorrelation method using clipping (AUTOC)
2) Cepstrum method (CEP)
3) Simplified inverse filtering technique (SIFT)
4) Data reduction method (DARD)
5) Parallel processing method (PPROC)
6) Spectral Equalization LPC method
7) Average magnitude difference function (AMDF)

However, among the above categories, the methods based on Autocorrelation and AMDF are more popular and widely used. A comparative study of each method is far from the goal in this paper. A detailed discussion on different approaches can be found in [4] and a comparative study between different methods has been done in [5]. Here since we use a modified version of AMDF in our work, we only have a brief overview on AMDF-based pitch detectors [6].

The AMDF pitch detection algorithm is chosen in our work because it has relatively low computational cost and is easy to implement. The principle of the pitch detection for speech signals using AMDF is based on the short-term difference function between each frame of speech signal and its lagged version, which is supposed to have a minimum when the lag is equal to pitch period:

$$AMDF_n(\eta) = \frac{1}{N}\sum_{i=1}^{N}|x_n(i) - x_n(i+\eta)|, \ \text{MinL} \le \eta \le \text{MaxL} \quad (1)$$

where $x_n(i)$ is n^{th} frame of the speech signal and MinL and MaxL is the minimum and maximum lags used for calculating of AMDF values for each frame, respectively.

Fig. 2 shows the result of applying AMDF for a periodic and an aperiodic signal. As it is seen in this figure, for a periodic signal, the period can be easily found by using AMDF. The difference function is expected to have a strong local minimum if in Eq. (1), η is equal to or very close to the period of the signal.

Based on this fact, AMDF is used for speech waveforms to measure the pitch period for voiced signals[1] which are semiperiodic. For each frame of speech signal, the lag where the AMDF is a global minimum is a strong candidate for the pitch period of that frame.

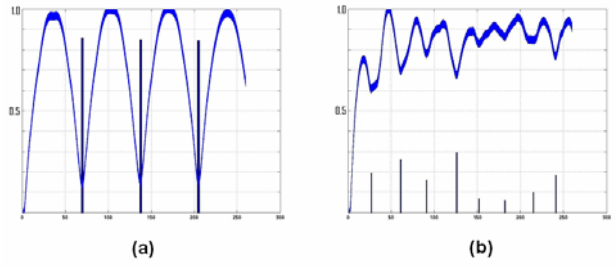

Fig. 2 AMDF for (a) a periodic and (b) an aperiodic signal.

In addition to the classical methods mentioned above, there has been also some efforts on using neural networks for the purpose of pitch detection [7], however the performance of most of these methods is related to the accuracy of a usually time-consuming preprocessing (for example peak detection), which may be also very sensitive to noise.

[1] Pitch period is not defined for unvoiced signals because there is no periodical excitation for unvoiced signals.

3. PROPOSED TWO-LEVEL PITCH DETECTOR

Fig. 3 illustrates the block diagram of the proposed approach for pitch detection which is done in a two-level hierarchy. We will show that such a two-level detection approach works well not only for a clean pitched signal, but for a noisy speech signals, as well.

As shown in this figure, the process starts with a preprocessing consisting of windowing and downsampling (under Nyquist condition). A cepstrum sequence[2] is then calculated for each frame of the speech waveform, which is used later in the first level of pitch detection to perform a voiced/unvoiced decision. In the second level, the pitch period extracted for the voiced signals using a modified version of a standard AMDF-based pitch detector. Such separation of voiced and unvoiced frames in the first level helps us to take advantage of the AMDF-based pitch detector which works quite well with voiced frames.

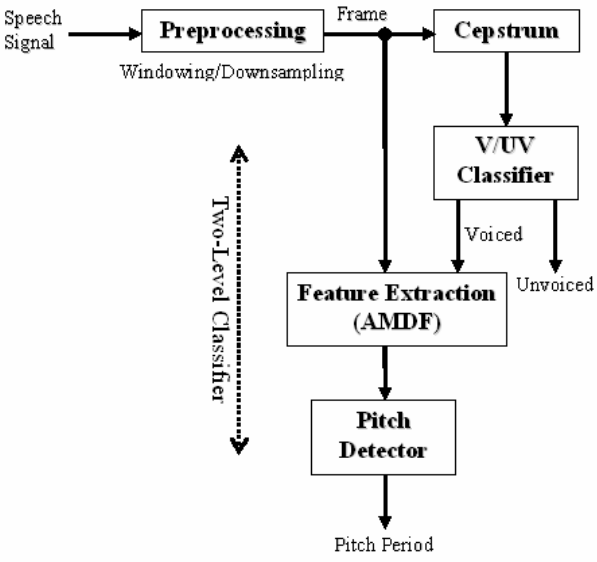

Fig. 3 Block diagram of the two-level proposed approach.

3.1 Short-Term Processing of Speech Signal

In general, since the long-term concept of a speech signal is not stationary, for extracting different features from speech, we usually work with short terms or frames of speech signal which are supposed to be stationary at each frame. For this purpose, we first select the desired *N*-length frame of the speech signal using a window, and then apply the short-term operation on this *N*-Length frame.

In this work, we used a Hamming window which is a common selection for the short-term processing, with a length of 200 samples that is well enough greater than the maximum possible pitch period in speech signals with a sampling rate of 8 kHz and short enough to reasonably satisfy the stationary condition for each frame.

To prevent from loss of information at the beginning and the end part of each frame, usually frames are extracted in an overlapping manner. It means that if the current frame is taken from the m^{th} sample of speech, the next frame starts from $m+L$ where L is less than the length of window. Here, we consider N=200 and L=100.

[2] Cepstrum is the Fourier transform of the logarithm of the spectrum of the speech signal which is known to be useful for determining periodicity in the spectrum [2] [3].

3.2 Voiced/Unvoiced Classification

Usually, misclassification of unvoiced frames as voiced and vice versa is a big problem in most pitch detectors and causes a lower performance especially when the signal is noisy, such problem is much more severe. To reduce such effects, we separate voiced/unvoiced decision from pitch period detection.

Based on this fact and also since the computation time is very important for real time applications, we use an MLP neural network [8] which in addition to its capability for parallel processing, is a powerfully well-developed classifier for nonlinear problems and if trained well, can discriminate almost precisely voiced from unvoiced signals even for speech in noisy environments that are usually the case in practice.

3.3 Modified AMDF Pitch Detector

Having discriminated the voiced frames, we are now ready to determine the pitch period for each frame in the second level using the AMDF pitch detection algorithm.

However, there are some problems that may happen when we use a standard AMDF-based pitch detector. Firstly, there is a possibility for detecting of a multiple of pitch period (e.g. double, triple etc. period) known as pitch period doubling, tripling and so forth, and secondly, the false voiced/unvoiced detection is usually higher in AMDF-based pitch detector in compare with other algorithms [5].

In this work, since we discriminate voiced from unvoiced signals in the previous step, the latter problem is not a matter of consideration any more. To overcome the pitch doubling problem, we did some modification on the AMDF algorithm for determining pitch period.

As mentioned in the previous section, in general the lag where the AMDF is a global minimum for a given frame of speech signal is selected for the pitch period of that frame:

$$P_i = \min_{\text{MinL}\le\eta\le\text{MaxL}}(AMDF_i(\eta)) \qquad (2)$$

where P_i is the value of pitch period for the i^{th} frame of the speech signal. The pitch values for all voiced frames will then make the pitch vector **P**. To correct the doubling values, we perform the following procedure:

1. Calculate mean m_p and standard deviations σ_p for **P**.
2. Delete $\mathbf{P}(j)<m_p-\sigma_p$ and $\mathbf{P}(j)>m_p+\sigma_p$, $1\le j\le \text{MaxF}$, where MaxF is the total number of frames.
3. Substitute deleted **P**(j)'s with:

$$\begin{cases}\text{next acceptable value}, & j=1\\ \text{previous acceptable value}, & j=\text{MaxF}\\ \text{average of previous and next accetable values}, & \text{otherwise}\end{cases}$$

3.4 Example of Pitch Detection

We give an example here to clarify the discussion. Fig. 4(a) and 4(b) illustrate a part of a speech signal with a sampling rate of 8 kHz and its corresponding frames. As mentioned before, the first level of the proposed approach makes a voiced/unvoiced decision based on the cepstrum sequence of each frame. The result of this process is shown in Fig. 4(c).

AMDF feature is then calculated for the voiced frames and the pitch value is selected for each frame based on the algorithm discussed in Section 3.3. Fig. 4(d) illustrates the calculated pitch period values (solid lines) as well as the actual value (dotted lines). As shown in this figure, except one voiced frame which detected as unvoiced, the detected pitch period value is almost identical to actual one.

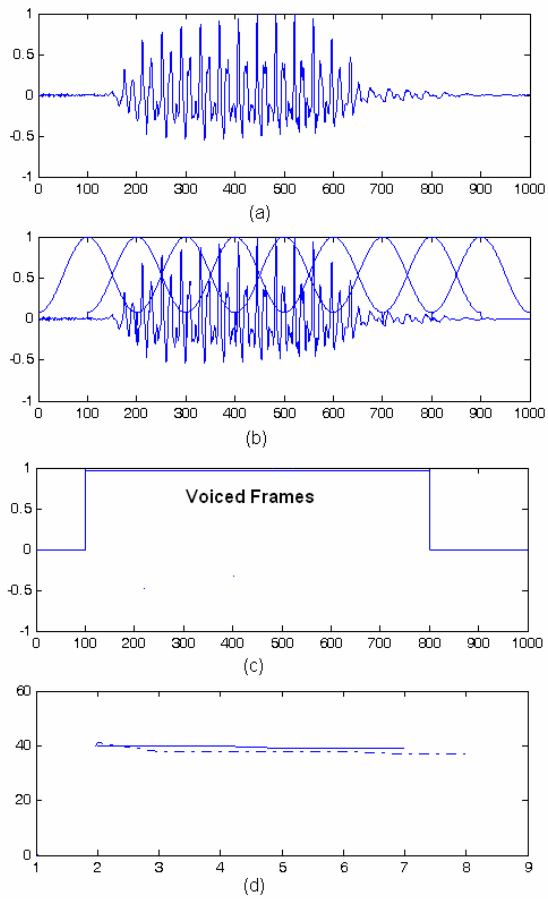

Fig. 4 A typical example of pitch detection. (a) speech signal, (b) extracted frames, (c) voiced/unvoiced decision, and (d) pitch period (solid line: calculated and dotted line: actual).

4. EXPERIMENTAL RESULTS

To train the MLP classifier and evaluate the accuracy of the pitch detector, we used two utterances from 6 different speakers (3 males and 3 females) drawn from the TIMIT database, a standardized corpus designed for acoustic phonetic research [9]. We extracted 1000 noise free frames from the waveform samples, including 900 voiced and 100 unvoiced frames. We then generated 2000 more frames which were the noisy versions of these samples using the following formula:

$$f_N(\text{n}) = f(n) + 0.1*r(n) \qquad (3)$$

where $f(n)$ is the frame signal and $r(n)$ is a random signal. Both $f(n)$ and $r(n)$ are normalized between -1 and 1.

Among 3000 existing frames, about 70% of the frames were used for training of MLP classifier which consists of one hidden layer with 20 neurons, 200 neurons in the input layer, and one output neuron which discriminates between voiced and unvoiced frames.

The remaining 30% data has been used for evaluation of the proposed pitch detector. For evaluation, we consider the following possibilities (Suppose that $p_s(m)$ as reference pitch values and $p_i(m)$ as detected pitch values).

i) $p_s(m) = 0$ and $p_j(m) = 0$ where both the standard analysis and the pitch detector classified the m^{th} interval as unvoiced. No error results here.

ii) $p_s(m) = 0$ and $p_j(m) \neq 0$ where the standard analysis classified the m^{th} interval as unvoiced, but the pitch detector classified as voiced. Here, an unvoiced-to-voiced error results.

iii) $p_s(m) \neq 0$ and $p_j(m) = 0$ where the standard analysis classified the m^{th} interval as voiced, but the pitch detector classified as unvoiced. Here, a voiced-to-unvoiced error results.

iv) $p_s(m) = P_1 \neq 0$ and $p_j(m) = P_2 \neq 0$ where both the standard analysis and the pitch detector classified the m^{th} interval as voiced. For this case, two types of errors can exist. If we define the voiced error e(m) as:

$$e(m) = P_1 - P_2 \quad (3)$$

Then, if $|e(m)| \geq 10$ samples (i.e., more than 1-ms error in estimating the pitch period), the error was classified as a gross pitch period error. For such cases, the pitch detector has failed dramatically in estimating the pitch period. The second type of pitch error was the fine pitch period error in which case if $|e(m)| < 10$ samples. For such cases the pitch detector has been supposed to estimate the pitch period sufficiently accurately.

Regarding the above points, we calculate four different types of error for seven standard detectors mentioned in Section 2 as well as our proposed framework. The errors are:

- Gross Error
- Voiced to Unvoiced (V/UV) Error
- Unvoiced to Voiced (UV/V) Error
- Deviation from Correct Pitch value

The results show that except for deviation from correct samples in fine detection which is still negligible, the performance of the proposed system is better than the other standard pitch detection algorithms for speech signals in the noisy environment. The improvement of the pitch detector can be specially recognized in V/UV and UV/V errors. Table 1 summarizes the results for the different pitch detectors.

Table 1 The evaluation results for different pitch detectors.

	Gross Error	V/UV Error	UV/V Error	Deviation (samples)
AUTOC	10 %	12 %	3 %	0.92
CEP	7.5 %	15 %	2 %	0.99
SIFT	8.5 %	6.5 %	7.8 %	0.91
DARD	13.5 %	7 %	5.7%	1.05
PPROC	19 %	6 %	7.3 %	1.09
LPC	14 %	5.5 %	32 %	0.88
AMDF	28 %	6.5 %	16 %	1.51
Proposed Approach	**7 %**	**6 %**	**2.5 %**	**1.51**

5. SPEAKER VERIFICATION

As mentioned in the beginning of the paper, our final goal is speaker verification in robot control system using the pitch period feature of the speech signal. A typical block diagram for such a purpose is illustrated in Fig. 5. As shown in this figure, in addition to the pitch information detected by the proposed approach, we might also extract the spectrum envelope of the speech signal (or any other related feature), and then characterize each authorized person for commanding based on these information.

Here, as an example of speaker verification, we evaluated such a typical system for this purpose, however to simplify the process, we have just considered the pitch information of speech signals. So, in practical case, if we consider the spectrum envelope of the signal too, the results will be better than the current results in this experiment.

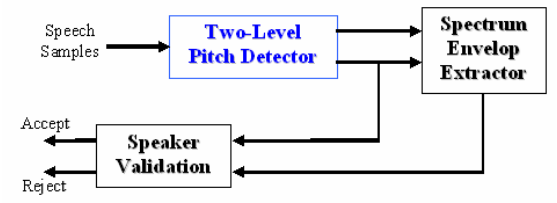

Fig. 5 using the proposed approach for speaker verification.

For this goal, we evaluated the system for 10 different commands used for robot control with 5 speakers (2 speakers as authorized and 3 others as unauthorized persons for commanding). The commands were: GO, STOP, LEFT, RIGHT, OPEN, CLOSE, PUSH, PULL, READ, and MUSIC. The results show that except for 3 cases out of totally 50 commands, the system successfully verifies the speakers.

6. CONCLUSION

In this paper, we presented a two-level pitch detector for natural speech signals which in compare with conventional pitch detectors has a better performance in the presence of noise. In the first level of pitch detection, we applied the cepstrum sequence of the speech to discriminate voiced from unvoiced signals by using a neural classifier. We then determined the pitch period value for the extracted voiced signals in the second level using a modified version of a typical AMDF-based pitch detection algorithm. Finally, we applied the proposed approach for the goal of speaker verification based on natural speech in a human-robot controller to validate the authorized speakers for commanding.

ACKNOWLEDGMENTS

This work was supported in part by GIST, in part by MIC through RBRC, and in part by MOE through BK21 project.

REFERENCES

[1] T. Fong, C. Thorpe, C. Baur, "Collaboration, Dialogue and Human-Robot Interaction," Proc. of 10th International Symposium of Robotics Research, Lorne, Victoria, Australia, November 2001.

[2] T. F. Quatieri, "Discrete-Time Speech Signal Processing: Principles and Practice," Printice Hall Signal Processing Series, ISBN: 013242942-X, 2002.

[3] J. R. Dellar (Jr), J. H. L. Hansen, J.G. Proakis, "Discrete-Time Processing of Speech signals," IEEE Press, ISBN: 0780353862, 2001.

[4] W. Hess, "Pitch Determination of Speech Signals: Algorithms and Devices," Springer-Verlag, 1983.

[5] L. R. Rabiner, A. E. Rosenberg, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. on Acoustics, Speech, and Signal Processing, 24:1 (399-418), 1976.

[6] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. on Acoustics, Speech, and Signal Processing, 22 (353-362), 1974.

[7] E. Barnard, A.R. Cole, M. Vea, and F. Alleva, "Pitch detection with a neural net classifier," IEEE Trans. on Acoustics, Speech, and Signal Processing, 39:2 (298-307), 1991.

[8] Y. H. Hu and J.N. Hwang, "Handbook of Neural Network Signal Processing," CRC Press, ISBN: 0849323592, 2002.

[9] DARPA TIMIT Acoustic-Phonetic Speech Database.