Sampling Rate Impact on the Performance of a Head-Related Impulse Response Decomposition Method

KENNETH JOHN FALLER II AND ARMANDO BARRETO
Electrical and Computer Engineering Department
Florida International University
Miami, FL 33174
USA
http://dsplab.eng.fiu.edu/

Abstract: - In order to achieve highly accurate 3D audio spatialization, an intended listener must undergo measurements in a specialized instrumentation system to obtain "individual" Head-Related Impulse Responses (HRIRs). Our goal is to create a customizable sound spatialization system that would not require complex individual measurements. Our approach is based on the reliable decomposition of measured HRIRs into damped sinusoidals. A previously developed method showed good performance in decomposing HRIRs collected by us at a sampling frequency of 96 kHz, but other HRIR databases comprise HRIRs recorded at lower rates. This paper compares the performance of our automated decomposition method with sampling rates of 96 kHz and 48 kHz.

Key-Words: - Head-Related Transfer Functions (HRTF), spatial audio, signal decomposition, damped sinusoidals

1 Introduction

Humans have the remarkable ability to localize sound in a three dimensional physical space. Many attempts have been made to create synthetic sounds that would cause humans to perceive a sound as if it were emanating from a source placed in a desired virtual 3D position, at a desired azimuth (θ), elevation (Φ) and distance (r) (Figure 1).

In the physical world our brains are capable to identify the location of a sound source thanks to binaural cues, such as the difference in time and intensity with which a sound will reach our eardrums, called interaural time difference (ITD) and interaural intensity difference (IID), respectively. However, sound localization is also determined by the specific way in which a sound is transformed as it travels from its source to each of our eardrums. These transformations can be modeled by linear transfer functions that mediate between the sound signal at its source and the sound delivered to each eardrum. Since it has long been recognized that the specific transfer functions are strongly influenced by the anatomical features of the listener, these are called "Head-Related Transfer Functions" (HRTFs), and their associated impulse responses are called "Head-Related Impulse Responses" (HRIRs). The spectral shaping imposed by each of these HRTFs on sounds originated at each position around the listener provides additional monaural cues for localization that are believed to be critical for elevation assessment. Furthermore, it is believed that much of that spectral shaping, such as the implementation of significant spectral notches and the potential for resonant effects, are closely related to the structural characteristics of the *pinna* or outer ear. Figure 1, shows a schematic depiction of the roles played by the Left HRTF (L-HRTF) and the Right HRTF (R-HRTF) in modeling the transformation of sounds that serves as the basis for our ability to localize sounds.

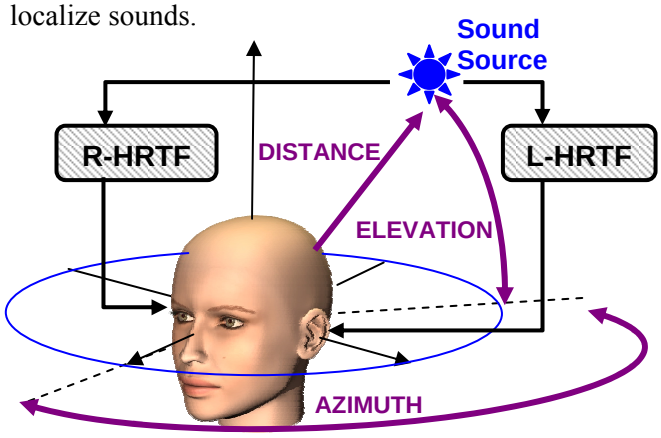

Fig. 1: HRTF modeling of sound transformation from a source to each of the listener's eardrums.

Several leading approaches to sound spatialization are based on the concept depicted in Figure 1. According to this model, if the dynamics of a given HRTF pair can be emulated with digital filters, then a

pair of binaural (Left and Right) sounds can be created filtering a monaural "source sound" through the R-HRTF and L-HRTF models, prompting the listener to perceive the sound as if emanating from the position for which the HRTFs were estimated. Furthermore, the HRIRs can be measured using dedicated equipment (e.g., "HeadZap" System, AuSIM, Palo Alto, CA) in which a compact sound source issues a signal that acts as a replacement for an impulse (e.g., Golay Codes), and the resulting sounds are captured and recorded from miniature microphones inserted in the entrance to the ear canal of the test subject (Figure 2). Spatialization can then be accomplished convolving a digital sound signal with the HRIRs recorded as long series (e.g., 256) of numerical values.

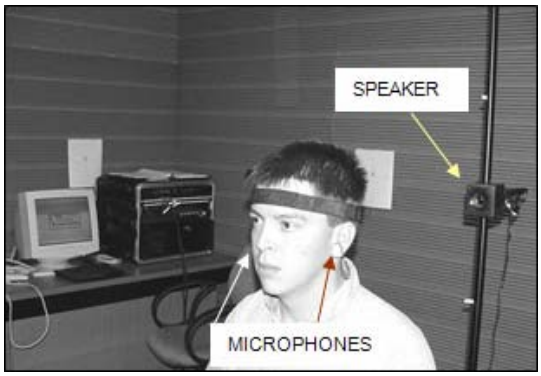

Fig. 2: Empirical HRIR measurement at FIU.

HRTFs and HRIRs are different for every location and differ from person to person. Ideally, each intended listener for an audio spatialization system should have his/her "individual" HRTFs (HRIRs) measured empirically, as described above. However, this is not practical for most audio spatialization applications. As a result, commercial developers have resorted to the use of "generic" HRIR pairs obtained experimentally from a mannequin of "average anatomical dimensions" (e.g., MIT's measurements of a KEMAR Dummy-Head Microphone [1]) or using a limited number of subjects to represent the general population (e.g., the CIPIC Database [2]). These databases include HRIR pairs for many different positions around the listener. Unfortunately, this type of "generic" HRIRs yield only an approximate sense of source location in many users, lacking the high spatialization fidelity of individual HRIRs [3].

The overall purpose of our research is to create a structural model for customizable HRIRs. Ultimately, the model would be customized by using the physical measurements of the intended listener to yield localization accuracy close to that of individually measured HRIRs. The current representation of HRIRs is complex and prohibits customization using the geometric characteristics of the intended listener. Therefore, we believe that decomposition of HRIRs into partial components will allow their re-generation from a reduced number of parameters that are related to the geometry of each intended listener. Efficient HRIR customization could have significant practical impact because it would extend the benefits of high-fidelity audio spatialization to the overall computer user population.

2 Methodology

The following subsections describe the methodology used in this study.

2.1 Structural Pinna Model

Previous attempts to develop structure-based HRTF models include Algazi's model for the effect of the listener's head, using only 3 simple anatomical measurements [4]. Brown and Duda's structural model in [5] accounted for the effects of the head, shoulder and the pinna features, which where cascaded together to generate a transfer function for the overall HRTF for each ear. This work, however, used ad-hoc assignment of the pinna model parameters, not based in anatomical features of the subjects

In [6], our group proposed a structural model for the effects of the pinna. This model comprised a resonator which feeds into one direct and three indirect paths. The indirect paths model reflected waves bouncing off the structures of the pinna before entering the ear canal. Recently, our group upgraded the model to include an additional (fourth) delayed wave (Figure 3).

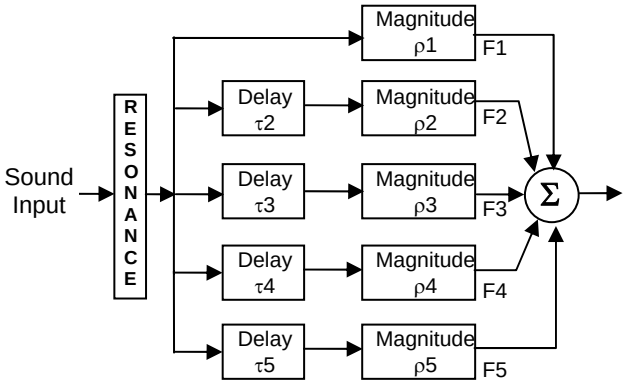

Fig. 3: Block diagram of the pinna model.

The parallel paths of the block diagram above represent the multiple signals entering the ear canal. As seen in the figure, each signal is delayed by a factor of τ_i and scaled by a factor of ρ_i. The pinna model shown in Figure 3 requires only a few parameters, which can be expected to relate to physical characteristics of the listener. Additionally, the model could be "cascaded" with Algazi's functional head model to represent a complete HRIR.

In order to instantiate the model show in Figure 3, an efficient method must be developed to obtain the parameters from empirically measured HRIRs. This will allow for the creation of a database of these parameters at various azimuths and elevations along with the measurements of the subjects' relevant anatomical characteristics. The database will then be used to develop an empirical relationship between model parameters and anatomical features. At that point the geometric characteristics of a new intended user could be measured and "converted" to parameter values to instantiate his/her model at a desired location.

2.2 Iterative Decomposition Method

The output of the model shown in Figure 3 will be a superposition of damped sinusoidals that appear at different delays and scales. If the HRIR is decomposed into these sinusoidal components, the delay and scaling parameters that should be used in the model would be revealed. In turn, these parameters can be used to create an instance of the HRIR model that will closely approximate the HRIR being decomposed.

Our group has developed a time domain method for decomposition of HRIRs [7, 8]. In this approach, two modeling methods, Prony and Steiglitz-McBride (STMCB), where compared for use in the decomposition of HRIRs. Both of these methods where used to model a second-order signal from a windowed portion of the HRIR in question. The window size had to be set so as to only contain a single damped sinusoid. A full description of this method can be found in [7, 8].

However, in that method the window sizes are not initially known. Hence, all possible window sizes must be iterated through to discover the appropriate sizes. The window sizes are gradually widened from 2 to 10 samples for each window, for a total of five windows. Under the assumption that an appropriately sized window would only contain a single damped sinusoid, one of the previously mentioned modeling methods could be applied to the samples contained within the window to fit the damped sinusoid present. Additionally, the window width eventually chosen would indicate the relative time delay of the next single damped sinusoid. The approximated single damped sinusoid would then be subtracted from the entire HRIR and the remainder of that subtraction is shifted to be left-justified in the analysis window. This process is repeated until five damped sinusoids and their delays are obtained.

After all potential damped sinusoids and their delays are extracted, they are summed together to obtain the candidate HRIR for that particular sequence of window widths. This candidate will be stored and the next sequence of window widths is explored until all possibilities are completed. Eventually, all the obtained candidates are compared to the original HRIR using Equations 1 and 2 and the highest fit is kept as the "Reconstructed" HRIR that represents the most accurate decomposition. Analysis of the results from this process showed that, in general, it approximates the original HRIR with relatively high accuracy. Figure 4 shows the components extracted from a measured HRIR by this process.

$$\text{Error} = \text{Original HRIR} - \text{Reconstructed HRIR}, \quad (1)$$

$$\text{Fit} = [1 - \{\text{MS(Error)}/\text{MS(Original HRIR)}\}]. \quad (2)$$

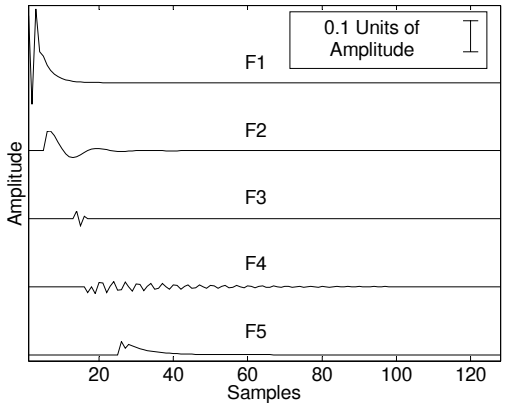

Fig. 4: Second-order HRIR components

In [8], process resulted in a 96% average fit but the iterative search was computationally expensive. An analysis of the search tree for just 5 window combinations revealed 9x9x9x9x9=59,049 leaf nodes that had to be explored. The addition of any other windows would multiply the leaf node total by 9 (per additional window). In order to select the best fit, all leaf nodes must be explored and the reconstructed

HRIR defined at that leaf node must be compared to the original HRIR to assess the fit. It quickly became apparent that this was not the most efficient mechanism for decomposition. It was also noticed that if the window size is small (less than 4 samples), STMCB and Prony will tend to inaccurately approximate the signal in question.

2.3 Decomposition via Pole Isolation

The previously described method depended critically on the assumption that there was some delay between the damped sinusoids that constitute an HRIR. While this assumption is, in general, valid, cases in which the delayed sinusoidals arrive very soon after each other will be particularly difficult to handle by this approach, given the limitations of the STMCB and Prony approximation methods mentioned above. Additionally, the window widths required by that method are not known in advance. This resulted in a search tree with a branching factor that remained high (e.g., 9) from the root node all the way to the leaf nodes.

In our new decomposition method, rather than windowing the segment to obtain some samples of the current sinusoid, the entire segment (at any point during the decomposition) is approximated by one of the modeling methods at a higher order. The candidate damped sinusoids for that particular stage of the decomposition are individually isolated according to their pole signature in the Z-domain.

In general, a single damped sinusoidal component sequence without a phase shift will be represented by a conjugate pair of poles within the unit circle and a zero at the origin of the Z-plane [9] (Figure 5).

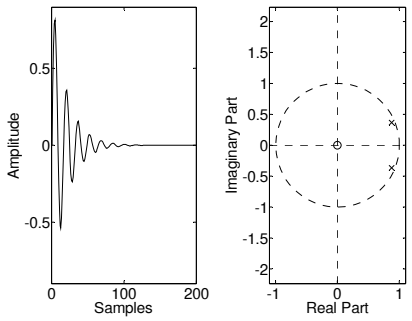

Fig. 5: Time domain and Zero-Pole plot representation of a single damped sinusoidal.

Hence, a damped sinusoid in the Z-domain can be described with the following general equation:

$$X(z) = \frac{k \cdot z}{(z - p_1)(z - p_2)} \tag{3}$$

where k is a scalar and p_1 and p_2 are complex poles. According to Equation 3, if the scalar k and the poles are known then, using the inverse Z-transform, it is possible to characterize the corresponding time domain sequence as a specific damped sinusoid.

As in the previous method, each candidate damped sinusoidal is isolated and subtracted from the current remnant of the HRIR. The delay factor is not predetermined, instead the remainder of the current segment is thresholded and the point at which the remnant surpasses the threshold is considered the point of delay. The remnant of the HRIR is then shifted, in time, to that point and re-approximated with one of the modeling methods but with an order which is two less. This process is repeated until five damped sinusoidals are extracted.

Similar to the previous iterative method, this method also results in a tree search. The advantage of this method is that amount of leaves decrease by one for each subsequent stage of the decomposition. If an analysis of the search tree is performed, one can see that only 5x4x3x2x1 = 5! = 120 leaf nodes need to be explored.

3 Pole Decomposition of HRIRs at 96 kHz and 48 kHz

The purpose of this study was to determine if the sampling rate of the measured HRIRs affect the performance of the pole decomposition method significantly. The decomposition method mentioned in the previous section was used to decompose HRIRs measured at Florida International University using the AuSIM HeadZap system at a sampling frequency of 96 kHz. For this study, this same set of HRIRs where down-sampled to 48 kHz and decomposed using the pole decomposition method as well.

The accuracy of the modeling process, as measured through our "fit" measure (Equations 1 and 2), varies according to the specific azimuth and elevation considered. As an example, consider Figures 6 and 7, which show the distribution of "fit" values for the azimuth and elevation pairs measured from subject XL's left ear, under both sampling rates. These figures demonstrate that the overall fit achieved for the majority of azimuth and elevation combinations is close to 0.9 (i.e., 90%), regardless of the sampling rate used, i.e., the ability of the method to find a

reasonable approximation of the HRIRs in question is not severely affected by the use of 48 kHz as sampling rate.

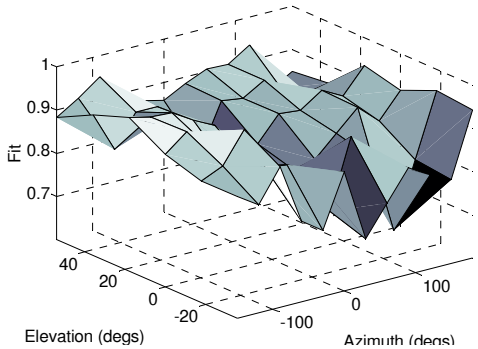

Fig. 6: Plot of elevation vs. azimuth vs. fit of the left ear of subject XL at 96 kHz.

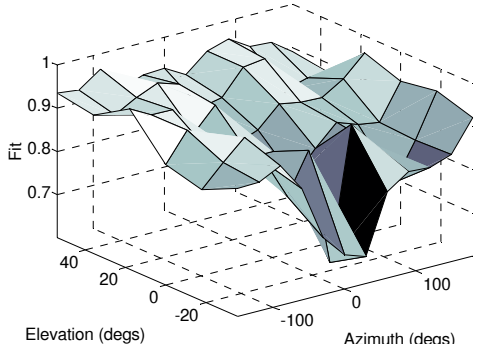

Fig. 7: Plot of elevation vs. azimuth vs. fit of the left ear of subject XL at 48 kHz.

Both these plots also indicated that the fits achieved for left-ear HRIRs were lower when the sound source was placed according to positive azimuths. Similarly, lower values of fit were recorded for very positive and, particularly, for very negative elevations.

In order to facilitate the comparison of the fit levels achieved under both sampling rates, Figure 8 shows the fits for all azimuth and elevation combinations in a two-dimensional graph. In this figure the fits obtained for a given azimuth and for all the measured elevations, i.e., a "slice" of Figure 6, are shown together. Then the set of measurements for the next azimuth and all measured elevations are shown, etc. Each label in the horizontal axis of this figure indicates where each one of these sets of measurements, sharing that azimuth, start (with an elevation of -36º). The following 5 points of the plot correspond to that same azimuth and elevations of -18º, 0º, 18º, 36º, and 54º, respectively. The figure shows the fits achieved at 96 kHz (solid line) and those achieved at 48 kHz (dashed line). Figure 9 shows the same type of display for the fits achieved in modeling the HRIRs from the right ear of subject XL.

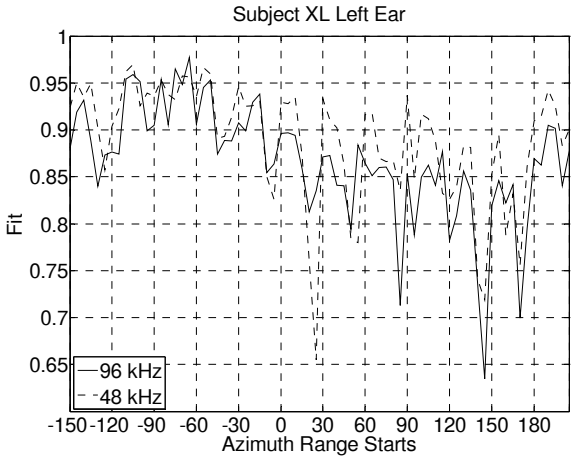

Fig. 8: Plot of the fits for the left ear of subject XL.

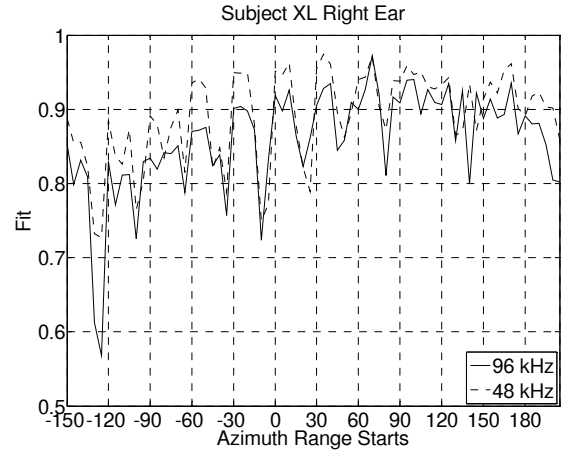

Fig. 9: Plot of the fits for the right ear of subject XL.

Figures 8 and 9 confirm the observations made from Figures 6 and 7 (lower fits for one hemisphere than the other and for extreme elevations) and, most importantly, show that fits achieved on HRIRs recorded at 96 kHz are not markedly different from those obtained on HRIRs studied at 48 kHz. Similar observations were made on results from 14 additional subjects.

It was of interest to examine the reconstructed HRIRs achieved for the highest and lowest fit cases, at both sampling rates. Figure 10 shows the highest fit case for XL's left ear, and Figure 11 shows the lowest fit case. It can be noticed that the use of 48 kHz or 96 kHz does not introduce a marked difference in the results.

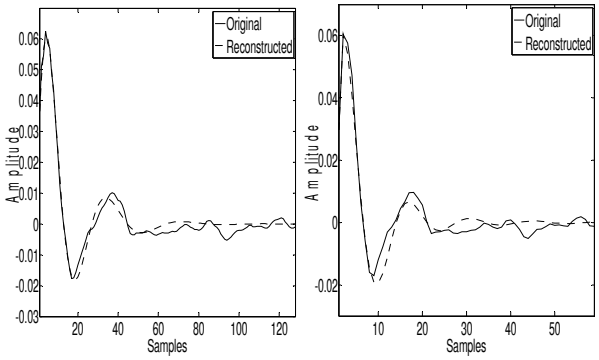

Fig. 10: Plot of original (solid) & reconstructed (dashed) HRIRs for subject XL's left ear at azimuth -90º , elevation -36º, using Fs=96 kHz (left panel) and Fs=48 kHz (right panel).

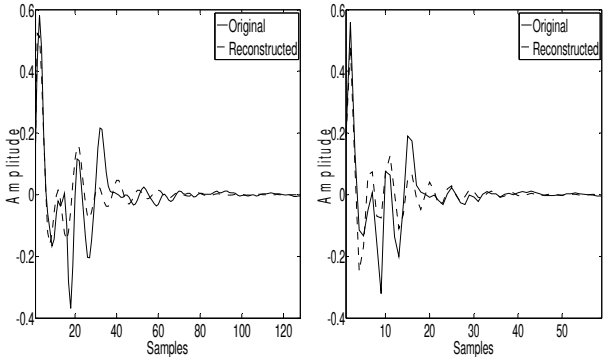

Fig. 11: Plot of original (solid) & reconstructed (dashed) HRIRs for subject XL's left ear at azimuth 120º , elevation -36º, using Fs=96 kHz (left panel) and Fs=48 kHz (right panel).

4 Conclusion

The results of using the proposed HRIR decomposition method on signals analyzed at a sampling rate of 48 kHz were very similar to the decomposition results achieved from the original signals collected at 96 kHz. Figures 8 and 9 show that the average levels of fit and also the variations of fit with azimuth and elevation follow the same patterns for both sampling rates. This is an important result, because it extends the feasibility of applying the decomposition method to HRIRs from publicly accessible databases, such as the CIPIC database [2], which, generally, are captured at 44.1 KHz.

5 Acknowledgments

This work was sponsored by NSF grants IIS-0308155, CNS-0520811, HRD-0317692 and CNS-0426125.

6 References

[1] B. Gardner, K. Martin, and Massachusetts Institute of Technology. Media Laboratory. Vision and Modeling Group., *HRFT measurements of a KEMAR dummy-head microphone*. Cambridge, Mass.: Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1994.

[2] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The Cipic HRTF database," in 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics New Paltz, NY, 2001

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer-Functions," *Journal of the Acoustical Society of America*, vol. 94, pp. 111-123, 1993.

[4] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of the Audio Engineering Society*, vol. 49, pp. 472-479, 2001.

[5] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *Ieee Transactions on Speech and Audio Processing*, vol. 6, pp. 476-488, 1998.

[6] A. Barreto and N. Gupta, "Dynamic Modeling of the Pinna for Audio Spatialization," *WSEAS Transactions on Acoustics and Music*, vol. 1, pp. 77-82, January 2004.

[7] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe, "Decomposition and Modeling of Head-Related Impulse Responses for Customized Spatial Audio," *WSEAS Transactions on Signal Processing*, vol. 1, pp. 354-361, 2005.

[8] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe, "Decomposition of Head Related Impulse Responses by Selection of Conjugate Pole Pairs," in International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 06), Bridgeport, CT, 2006

[9] J. Proakis and D. Manolakis, "Digital Signal Processing: Principles, Algorithms and Applications," 3rd ed. Upper Saddle River, NJ: Prentice Hall, 1995, pp. 174.