MUSIC SUMMARY USING KEY PHRASES

Stephen Chu Beth Logan

Cambridge Research Laboratory

Technical Report Series

CRL 2000/1
April 2000

COMPAQ

Cambridge Research Laboratory

The Cambridge Research Laboratory was founded in 1987 to advance the state of the art in both core computing and human-computer interaction, and to use the knowledge so gained to support the Company's corporate objectives. We believe this is best accomplished through interconnected pursuits in technology creation, advanced systems engineering, and business development. We are actively investigating scalable computing; mobile computing; vision-based human and scene sensing; speech interaction; computer-animated synthetic persona; intelligent information appliances; and the capture, coding, storage, indexing, retrieval, decoding, and rendering of multimedia data. We recognize and embrace a technology creation model which is characterized by three major phases:

Freedom: The lifeblood of the Laboratory comes from the observations and imaginations of our research staff. It is here that challenging research problems are uncovered (through discussions with customers, through interactions with others in the Corporation, through other professional interactions, through reading, and the like) or that new ideas are born. For any such problem or idea, this phase culminates in the nucleation of a project team around a well-articulated central research question and the outlining of a research plan.

Focus: Once a team is formed, we aggressively pursue the creation of new technology based on the plan. This may involve direct collaboration with other technical professionals inside and outside the Corporation. This phase culminates in the demonstrable creation of new technology which may take any of a number of forms—a journal article, a technical talk, a working prototype, a patent application, or some combination of these. The research team is typically augmented with other resident professionals—engineering and business development—who work as integral members of the core team to prepare preliminary plans for how best to leverage this new knowledge, either through internal transfer of technology or through other means.

Follow-through: We actively pursue taking the best technologies to the marketplace. For those opportunities which are not immediately transferred internally and where the team has identified a significant opportunity, the business development and engineering staff will lead early-stage commercial development, often in conjunction with members of the research staff. While the value to the Corporation of taking these new ideas to the market is clear, it also has a significant positive impact on our future research work by providing the means to understand intimately the problems and opportunities in the market and to more fully exercise our ideas and concepts in real-world settings.

Throughout this process, communicating our understanding is a critical part of what we do, and participating in the larger technical community—through the publication of refereed journal articles and the presentation of our ideas at conferences—is essential. Our technical report series supports and facilitates broad and early dissemination of our work. We welcome your feedback on its effectiveness.

Robert A. Iannucci, Ph.D.
Vice President, Corporate Research

Music Summary Using Key Phrases

Stephen Chu
University of Illinois
Urbana IL 61801

Beth Logan
Cambridge Research Laboratory
Cambridge MA 02139

April 2000

Abstract

As the magnitude and use of multimedia databases grows rapidly, efficient ways to automatically find the "gist" of the contents becomes a necessity. This work addresses the problem of summarizing music, specifically songs of rock or pop genre. We assert that a typical song can be summarized by one or more representative "key phrases". Our goal then is to identify the reoccurring temporal patterns in the audio signal.

We investigate two approaches to summarization. In the first, we divide a song into fixed-length segments and cluster these based on a similarity measure. The key phrase is extracted by choosing the most frequent cluster. In the second approach, Hidden Markov Models (HMMs) are used to discover the structure of the music. Both approaches use the Mel-frequency cepstrum as the feature.

The two proposed approaches were evaluated on a set of Beatles songs. The clustering method consistently achieved good performance, while the HMM method produced mixed results.

CRL Technical reports are available on the CRL's web page at
http://www.crl.research.digital.com.

Compaq Computer Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700, Suite 721
Cambridge, Massachusetts 02139
USA

1. INTRODUCTION

Automatic music summarization aims at extracting concise gist from music, so that interaction with large multimedia database can be made simpler and more efficient. Potential applications of automatic music summarization include multimedia indexing, multimedia data searching, content-based music retrieval, and online music distribution.

Research efforts have been reported in several related areas. In the area of music content analysis, Vercoe's group in the MIT media lab built a system to perform speech-music discrimination, beat and tempo tracking, and timbre classification [1]. Automatic music transcription has been intensively studied [2]-[4]. Although there are a number of well-understood time-domain and frequency-domain techniques for monophonic transcription (pitch tracking), only limited success is achieved for polyphonic music [4], [5]. In the area of content-based audio retrieval, sound segments are retrieved based on the similarity of the query template and the database entries. Experiments reported by Foote [6] as well as Wold *et al* [7] show good results on a data set containing short sound files. Recent work in music retrieval allows the user to hum or whistle a desired tune and interact with the database using the melodic query [8], [9]. Finally, in the area of multimedia indexing, various research groups have successfully demonstrated systems that summarize video clips using key-frames [10].

We assert that a typical song can be summarized by one or more representative "key phrases". Our goal then is to identify the reoccurring temporal patterns in the audio signal. Furthermore, this work addresses one of the fundamental problems in music content analysis: to discover the structure of a piece of music.

An intuitive if somewhat naïve approach to the problem is to first transcribe the song into score and then look for repetitive patterns or motifs in the melody. Unfortunately, as demonstrated in [5], to reliably find the melody in a complex arrangement is all but impossible using present technologies. Moreover, pattern discovery in temporal sequences is also a difficult problem, let alone the fact that the sequence is usually noisy. To avoid this problem, instead of the melody we use cepstral features, which capture the entire spectrum of the audio signal. Based on these features, we investigate two approaches to summarization. In the first, we divide a song into fixed-length segments and cluster these based on a similarity measure. The key phrase is extracted by choosing the most frequent cluster. In the second approach, Hidden Markov Models (HMMs) are used to discover the structure of the music. Given a song, unsupervised training is performed to generate an N-state ergodic HMM. We then use this HMM to get the Viterbi alignment of the song. Consecutive frames having the same states are joined to form continuous segments and the key phrases are chosen based on the duration and frequency of occurrence of these segments.

This report is organized as follows. In section 2, the calculation of cepstral features is briefly introduced. In addition, we provide further justification for considering the

spectra rather than melody. The proposed clustering approach and the HMM approach to music summary are discussed in section 3 and section 4 respectively. The experiments and results are shown in section 5. We finally conclude the report in Section 6.

2. ACOUSTIC FEATURES

As discussed in the previous section, automatic transcription of polyphonic music is a difficult task. In fact, it is also unnecessary for the purpose of discovering of structures in music [1]. According to Martin *et al*, the acoustic structure of music is more important than its written form. This is evident in human perception of music. Human listeners hear groups of notes, or chords, as single objects in many circumstances. Based on these findings, we argue that a good acoustic feature set for automatic music summary should cover a broad range of the spectrum. Furthermore, using parameterized spectra as our features allows us to capture timbre and energy changes, which often correspond to salient events in music such as arrangement transitions and carry important information about the structure of the song. An example of such a transition is shown in Figure 1.

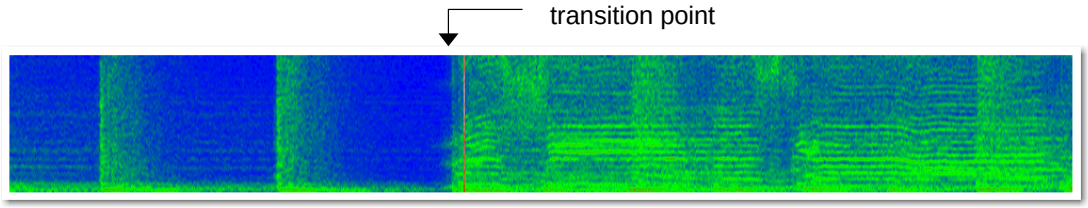

Figure 1. Structures embedded in the spectrum

2.1. Calculation of Cepstral Features

The cepstral coefficients provide a compact representation of the spectra and are widely used in speech processing [11]. The calculation of cepstral features is summarized in Figure 2.

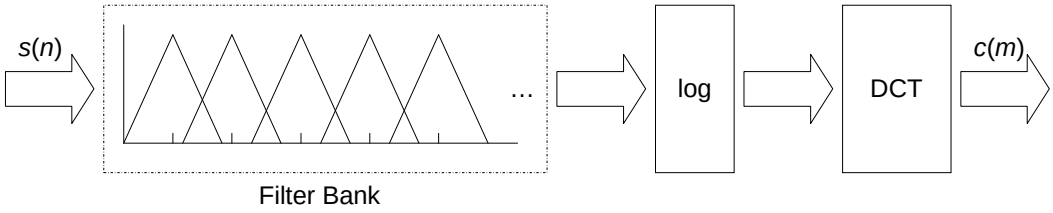

Figure 2. Computing cepstral coefficients

The sampled data is first passed through filter-bank analysis. The cepstral coefficients are calculated from the log filter-bank amplitudes using the discrete cosine transform (DCT). Psychophysical studies have shown that human perception of the audio

frequency content does not follow a linear scale. In our experiments, we use the warped mel-scale that is commonly used in speech recognition applications.

The cepstral coefficients give good discrimination and have properties that can be helpful to the music analysis task. In particular, the cepstrum is a decaying sequence, and can be truncated to obtain a smoothed estimate of the log spectra. We know that variability of the lower cepstral coefficients is primarily due to variations in the characteristics of the sound source. For speech recognition, these variations are considered as noise and are usually de-emphasized by cepstral weighting. However, when analyzing music, the differentiation of the generating source (stings, drums, vocals, etc.) becomes more important than the phonetic content. Therefore, we can benefit from this property by emphasizing the lower order cepstral coefficients.

3. CLUSTERING APPROACH TO MUSIC SUMMARY

The task of extracting key-phrases form music is, in essence, a pattern discovery problem. Although pattern discovery shares many similarities with the well-studied area of pattern recognition, it also poses a different challenge than the latter. For example, in the retrieval by melody experiments [8], [9], the tune hummed by the user is a defined target pattern used in recognition. However, in the case of key-phrase finding, there is no pre-defined pattern to look for. Nor do we posses any prior information about the location and duration of the pattern in the data stream. Hence, the underlying challenges are two-fold: 1. To get plausible segmentations of the sequence; and 2. To identify and extract patterns. Note that these two tasks don't necessarily need to be solved in separate steps.

The clustering approach tackles the problem in two steps. First, a given song is divided into fixed length segments. These segments are grouped into clusters based on their cross-entropy measures. In the second step, the clusters are sorted by their frequencies of occurrence. The longest example of the most frequent episode is chosen as the key-phrase.

3.1. Clustering Algorithm

The segments are clustered together based on second-order statistics, using the Cross-Entropy measure. The algorithm is implemented using the following iterative procedure.

1. Every member has its own cluster.

2. Compute and store the resultant distortion of combining any two clusters.

3. Pick the pair with the lowest distortion.

4. If it is less than a predefined threshold, then combine the two clusters and go to step 2.

5. If not, quit.

The algorithm is summarized in Figure 3.

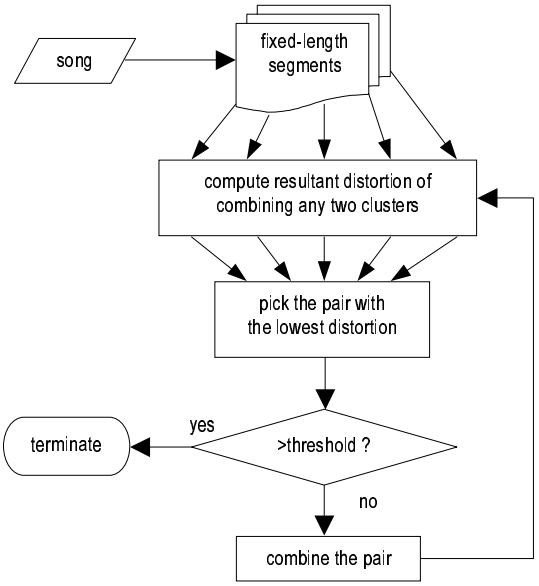

Figure 3. The clustering algorithm

3.2. Distortion Measure

The distortion measure is a modified cross-entropy or Kullback Leibler (*KL*) distance. A modification to the standard distance is needed to make it symmetric.

We use

$$KL2(A;B) = KL(A;B) + KL(B;A) \tag{1}$$

where *A* and *B* are two distributions and

$$KL(A;B) = E[\log(pdf(A)) - \log(pdf(B))] \tag{2}$$

Assuming the pdfs *A* and *B* are Gaussians, then

$$KL2(A;B) = \Sigma A/\Sigma B + \Sigma B/\Sigma A + (\mu A - \mu B)^2 \cdot (1/\Sigma A + 1/\Sigma B) \tag{3}$$

where $\Sigma *$ denotes variance and $\mu *$ denotes mean.

When not enough data points are available in a cluster, the Mahalanobis distance is used.

$$M(A;B) = (\mu A - \mu B)^2 / \Sigma A \quad (4)$$

Note that equation (4) is for the case where cluster *B* doesn't have enough points.

3.3. Limitations of the Clustering Approach

The algorithm has fixed resolution because of the initial set of clusters. This can result in quite unnatural segmentations. The effect of hard segmentation can be reduced by decreasing the segment size or by sliding a window in small steps over the data stream. However, due to its computationally expensive nature, the $O(n^2)$ clustering algorithm quickly exceeds the limit of computing power as the number of initial clusters increases. The HMM approach discussed in the next section provides an elegant solution of the problem.

4. THE HMM APPROACH

In the HMM approach, we seek a statistical model that reflects the structure of the data. Instead of arbitrarily dividing the data, we hope to "learn" the segmentation from the data itself. Moreover, the HMM method integrates the data segmentation process and pattern discovery into a unified process. In addition, there exist efficient algorithms for the training and decoding of HMMs.

4.1. Unsupervised Learning using HMM

The HMM is a powerful framework for learning and recognition of temporal patterns and has found great success in many pattern recognition applications such as speech recognition. In those cases, the data are usually assumed to belong to a number of different classes. An HMM is trained for each class based on labeled training examples believed to be from that class. During recognition, when an unknown utterance is presented, the HMM that is most likely to generate the given observations is selected.

In our case, however, no labeled training data is available and the classes are not defined. In fact, our goal is to find the "classes" and the underlying structure of the generating process. This is achieved by unsupervised learning of ergodic HMMs.

An ergodic HMM is a fully connected finite-state machine. One HMM is trained for a given song. The hope is that each state will correspond to a group of similar segments in the song. The Viterbi algorithm finds the most likely state sequence given the data and the model. By interpreting the Viterbi alignment results, we can infer the structure of the song. In the experiment, consecutive frames having the same states are joined to form

continuous segments and the key phrases are chosen based on the duration and frequency of occurrence of these segments. The process is illustrated in Figure 4.

The Baum-Welch algorithm is used for the training of the HMMs. The details of the Baum-Welch algorithm and the Viterbi alignment algorithm can be found in [11].

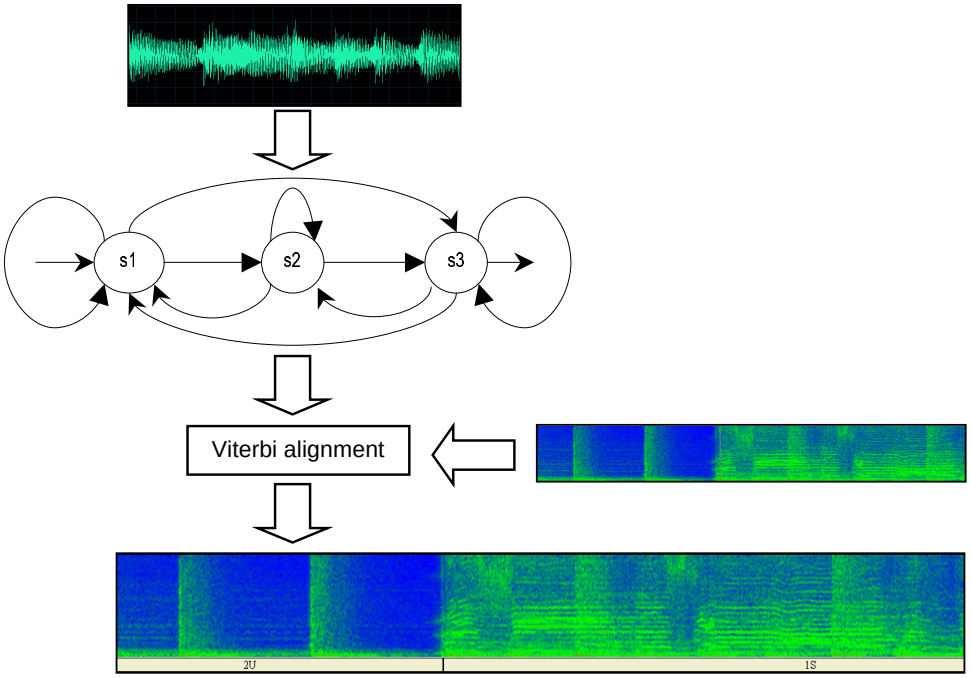

Figure 4. Discover structure using unsupervised learning

5. EXPERIMENTS AND RESULTS

Unlike speech recognition where the performance of a system can usually be precisely evaluated by comparing the recognition results with the text transcription, there is unfortunately no ground truth in automatic music summary. The only validation is through well-designed user tests.

5.1. Experiment Paradigm

In our experiment, three 10-second summaries are generated for each song: one from each of the two proposed methods and another cropped randomly from the song. A

human listener is presented with these three excerpts without the knowledge of the correspondence between the clips and the generating methods. For each of the excerpts, the user is asked to give a rating of "good", "average", or "bad". The user can choose to skip the song if he/she does not feel familiar with the song well enough to make a judgement. The songs are drawn from a song pool and presented in random order.

We confine the scope of the experiment to songs of rock or pop genre. In our preliminary tests, we found that it is easier for people to make confident and consistent judgements if they know the song. Therefore, we use a pool of 18 No.1 Beatles songs in the evaluation, assuming that these songs are familiar to the users. This is a relatively more objective way to define a test set comparing with picking the songs manually.

For both the clustering and the HMM method, we investigated a number of parameterization variations, namely, applying different analysis window sizes, using different analysis orders, introducing various band-limiting filters, and incorporating delta cepstral features. For the HMM approach, we also tried models with different number of states and different topologies. We would like to point out that these trials are not meant to be exhaustive, but to provide us with a reasonable starting point. The following are the finalized settings for the two methods used in the user evaluation.

The clustering method divides the data into non-overlapping 1-second segments. The HMM method uses a 0.9-sec sliding window with 0.1-sec frame rate. In both approaches, the audio in each analysis window is sampled at 16KHz and band-limited to 133Hz-6,855Hz. A mel-scaled filter bank consisting of 40 bandpass filters is used to calculate the cepstral coefficients. 13 cepstral coefficients are used in the clustering methods and 2 in the HMM approach. The HMM has three fully connected emitting states and one non-emitting exit state.

The Compaq *Calista* large vocabulary speech recognition package is used for the training of the HMMs. Necessary modifications were made to enable the handling of ergodic models. The training procedure was altered to perform Viterbi state alignment.

5.2. Results and Discussions

10 users participated in the evaluation. We enumerate the ratings "good", "average", and "bad" to 3, 2, and 1 respectively. The results are summarized in Figure 5.

From the figure, we can see that the clustering method achieved good performance, while the HMM method only did a mediocre job. It is not surprising that the random method did not fare poorly. This is because most pop songs are quite repetitive in nature, and the chance of hitting a "good" excerpt in a random 10 second segment is reasonably high. However, although the HMM method and the random method receive similar performance ratings, the former consistently starts a summary at more natural places (such as the beginning of a phrase) then the latter. In fact, the HMM method is also

capable of finding natural stopping points for the key-phrases. But in order to make the comparison fair, all summaries are chopped to 10 seconds.

Whether a summary is good or not is a subjective matter and the answer can vary noticeably among different users. The error bar on Figure 5 shows the standard deviation of scores from different users. It is observed that the HMM method is more "controversial" than the others. We did not discuss with the test subjects how they should evaluate the summaries before they perform the tests. However, after the tests we asked them what criteria they used. We found that users felt that: 1. A vocal portion is better than instrumental. 2. It's nice to have the title sung in the summary. 3. The beginning of the song is usually pretty good, at least that gets an average. 4. It's preferable to start at the beginning of a phrase rather than in the middle.

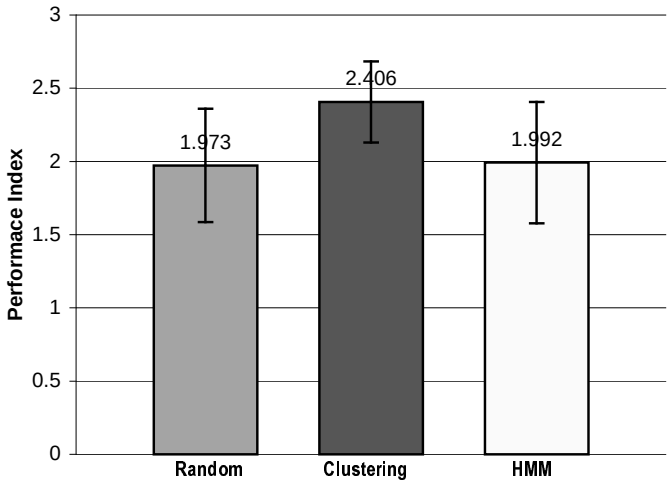

Figure 5. Relative performance of the methods (1.0:poor; 2.0:average; 3.0: good). Error bar indicates standard deviation across users.

We noticed that the summaries generated by the HMM method tend to receive more extreme ratings, e.g. either "good" or "bad", and relatively few "average". Examination of the bad clips reveals that many of these are pure instrumental sections. The HMM method is good at finding boundaries between sections with different spectral characteristics and can usually precisely detect the transition between a vocal and instrumental part. However, the method does not have the capability of classifying whether a segment is vocal or not. Potentially, the performance of the HMM method can be greatly improved by incorporating explicit singing detection.

6. CONCLUSIONS AND FUTURE WORK

This work addresses the problem of automatic music summary. We investigated two approaches to finding key-phrases based on clustering segments and learning HMM structure. Subjective experiments on Beatles songs show that the clustering approach

gave consistent better than random summaries. The HMM method is capable of uncovering structures in music but was not better than random.

The structure discovery potential of the HMM approach is limited by the inflexibility of a fixed model topology. Recent work on entropy minimization by Brand shows promising results in pattern discovery [12]. Also, for the songs studied, an instrumental rather than preferred vocal summarization was often selected. Future work would therefore involve investigation of these two areas.

REFERENCES

[1] K. D. Martin, E. D. Scheirer, and B. L. Vercoe, "Music Content Analysis through Models of Audition," *Proc. ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*, Bristol UK, Sept. 1998.

[2] J. Brown, "Musical Fundamental Frequency Tracking Using a Pattern Recognition Method," *J. Acoustical Soc. of Am.*, vol. 92, no. 3, pp. 1394-1402, 1992.

[3] J. Brown and B. Zhang, "Musical Frequency Tracking Using the Methods of Conventional and 'Narrowed' autocorrelation," *J. Acoustical Soc. of Am.*, vol. 89, no. 5, pp. 2346-2354, 1991.

[4] L. Grubb and R. Dannenberg, "A Stochastic Method of Tracking a Vocal Performer," *Proc. Int'l Computer Music Conf.*, pp. 301-308, 1997.

[5] K. D. Martin, "Automatic Transcription of Simple Polyphonic Music," *Tech Report*, MIT Media Laboratory Perceptual Computing Section, No. 309, 1996

[6] J. T. Foote, "Content-Based Retrieval of Music and Audio," in C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, vol. 3229, pp. 138-147, 1997.

[7] E. Wold, T. Blum, D. Keslar, and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, pp. 27-36, Fall 1996.

[8] A. Ghias et al., "Query by Humming," in *Proc. ACM Multimedia 95*, San Francisco, Nov. 1995.

[9] R. McNab, L. Smith, I. Witten, C. Henderson, and S. Cunningham, "Towards the Digital Music Library: Tune Retrieval From Acoustic Input," in *Proc. Digital Libraries '96*, pp. 11-18, 1996.

[10] Y. Zhuang, Y. Rui, T. S. Huang, "Video Key-Frame Extraction by Unsupervised Clustering And Feedback Adjustment," *Journal of Computer Science & Technology*, vol. 14, no. 3, pp. 283-287, 1999.

[11] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

[12] M. Brand, "Structure Discovery in Conditional Probability Models via an Entropic Prior and Parameter Extinction," *Neural Computation*, July 1999.

CRL 2000/1
April 2000

MUSIC SUMMARY USING KEY PHRASES

Stephen Chu and Beth Logan

COMPAQ