THE SHAZAM MUSIC RECOGNITION SERVICE

Guided by a user's query-by-example music sample, it delivers the matching song, as well as related music information, of immediate interest to the user.

By Avery Wang

People are routinely exposed to music in everyday environments—car, home, restaurant, movie theater, shopping mall—but are frustrated by not being able to learn more about what they hear. They may, for instance, be interested in a particular piece of music and want to know its title and the name of the artist who created it. They may want to buy a digital download of the song or a ringtone. To address these limitations, my colleagues and I at Shazam Entertainment have developed a query-by-example (QBE) music search service that enables users to learn the identity of audible prerecorded music by sampling a few seconds of audio using a mobile phone as a recording device.

TVision Insights, Inc. – Ex. 1026, Page 000001
IPR2023-01014 (TVision Insights, Inc. v. The Nielsen Company (US), LLC)

Others have also sought to deliver music identification. For example, in 1999 an early pioneer, StarCD, introduced a service enabling users with mobile phones to identify songs playing on certain radio stations [1]. Identification was accomplished by users calling into the StarCD service while an unknown song was playing. They would then enter the radio station call letters onto the phone's keypad. StarCD would automatically consult a third-party playlist database provider to determine what song was playing on the specified radio station at the time of the call. Learning the identity of the music, users would be given the opportunity to buy the CD. The StarCD system was thus limited to songs playing on radio stations being monitored by the third-party playlist provider.

Shazam interface: (a) recording a 10-second audio sample; (b) displaying the query results.

QBE is a more flexible music-recognition modality than the one provided by StarCD, recognizing music from a sample recording rather than requiring the user to key in radio station information. QBE music recognition research has been pursued by a number of groups worldwide. For example, in the early 1980s Broadcast Data Systems (www.bdsonline.com/) was an early pioneer, using a variety of methods to correlate waveforms [4]. In 1996 Musclefish (www.musclefish.com/) developed a method based on multidimensional feature analysis and Euclidean distance metrics [6]. However, these methods were all best suited for fairly clean audio samples and do not work well in the presence of significant noise and distortion.

When we founded Shazam Entertainment in London in 2000, we aimed to develop a commercial QBE music recognition service that could be offered through mobile phones [5]. A user would capture audio samples by calling the service through a short dial code; after sampling, the server

TVision Insights, Inc. – Ex. 1026, Page 000002

THE ADDITION OF MANY millions of songs into the database must not significantly decrease the probability of finding a correct match or significantly increase the probability of reporting a false positive.

would hang up and return the identification results via an SMS text message. In 2004, we introduced a newer version called Song Identity, providing a more interactive interface, available first through Verizon Wireless in the U.S. Since then, we have ported the application to a variety of handset platforms.

Before using Song Identity, users first download and install its applet into their handsets. Upon hearing music they wish to identify, they launch the applet, which records about 10 seconds of ambient audio and either sends an audio file to Shazam or performs feature extraction on the phone to generate a small signature file. This signature file is then sent to a central server that performs a search and provides the matching metadata—title, artist, album—back to the applet (see the figure here). Users are then presented with the title, artist, and album information, as well as the option of downloading the corresponding ringtone—if the matching one is available—based on the music. Future versions on some platforms may also allow the purchase of full-track downloads, CD, concert tickets, lyrics, and other follow-on data or activities associated with the music.

Commercial QBE music recognition systems, including Song Identity, must overcome a number of daunting technical challenges:

Noise. Noise competes with the target music in some environments; for example, music is generally in the background in noisy cafés and shopping malls. Being in a car adds the complication of traffic or engine noise. Noise power may indeed be significantly greater than signal power, so a recognition algorithm must be robust enough to deal with significant noise.

Distortion. The system must be able to deal with distortion arising from a variety of sources, including imperfect playback or sampling equipment, as well as environmental factors (such as reverberation and absorption). Sampling through telephony equipment reduces the frequency response to about 300Hz–3,400Hz. Distortion may also arise from the audio sample being subject to low bit-rate voice compression in the mobile phone. Nonlinear noise suppression and voice-quality-enhancement algorithms built into the handset or into the mobile carrier's network may represent a further challenge in which the sampling of background music results in recordings that contain mostly silence.

Database management. The system must be able to index "fingerprints" of millions of songs in its online database without requiring an inordinate number of servers. Thus the fingerprint, or unique feature representation, of each song must be reasonably small, on the order of only a few kilobytes. Moreover, scaling to millions of songs must not be able to incur a significant processing load on the backend search engine, as the system may need to dispatch hundreds or thousands of queries per second. The system must also scale statistically, meaning the addition of many millions of songs into the database must not significantly decrease the probability of finding a correct match or significantly increase the probability of reporting a false positive.

We were dismayed by the early audio samples we collected through mobile phones; the music was often so distorted we could barely recognize the presence of music with our own ears. We were often hard-pressed to match an audio sample over the phone against its known master recording, let alone scale it to millions

TVision Insights, Inc. – Ex. 1026, Page 000003

of recordings. We were even at risk of having to abandon our efforts. Fortunately, in 2000 after three months of work, we arrived at a solution involving temporally aligned combinatorial hashing, generally overcoming these challenges.

Another challenge we managed to address is more logistical than technological: How to cost-effectively compile a database of millions of songs. Shazam has been purchasing music assets, as well as extracting fingerprints from content partners with large catalogs of music. Confronting these constraints actually helped simplify development of our music recognition algorithm. We were forced to disregard approaches that could not scale to large numbers of recordings, especially in the presence of more noise than signal. This thinking produced a number of insights [5]; first among them, we would have to find robust features that could be reproduced in the presence of significant noise and distortion. We considered a number of candidate features (such as power envelopes and mel-frequency cepstral coefficients), but most weren't robust enough for our needs.

We needed features that could be linearly superposed (transparently) and recovered in the presence of noise. We thus turned to spectrogram peaks [1], which provide a map of energy distribution in terms of time and frequency.[1] The location of the peaks, though the result of nonlinear processing, are substantially linearly superposable; that is, a spectrogram peak analysis of a mixture of music and noise contains spectral peaks (due to the music and the noise) if each would be analyzed separately. The presence of corresponding spectrogram peaks in a noisy versus noiseless music signal makes it possible to determine with high probability whether an audio sample matches a recording in the database.

Though robust, sets of individual spectrogram peaks used as fingerprint features provide insufficient entropy (the number of unique features is too small) to allow for an efficient search, especially in a very large database. In order to increase the entropy while maintaining transparency, we hit upon a scheme we call "combinatorial hashing" in which we construct fingerprint hash tokens using pairs of spectrogram peaks chosen from the set of all spectrogram peaks present in the signal being analyzed. In the fingerprint formation process, we use a subset of spectrogram peaks as "anchor points," each with a target zone defined by a range of time and frequency values offset from the anchor point's coordinates. Each anchor point is paired to a number of target points in the target zone. The frequency information and relative time offset from each pair of points are used to create a 32b fingerprint hash token.

This combinatorial expansion results in perhaps a ten-fold increase in the number of tokens searched in the database over the original number of spectrogram peaks. However, the increased entropy in the hash tokens helps accelerate the index search by a factor of more than a million, resulting in significant speed improvement when identifying a particular song. This speedup is due to the fact that more descriptive bits of information allow for cutting more efficiently through ambiguous clutter.

The effect of this combinatorial hashing on a mixture of music and noise generates three classes of fingerprint tokens:

- Both spectrogram peaks belong to the target signal;
- One peak belongs to the target signal and one to a noise signal; and
- Both peaks belong to noise.

Only the tokens having peaks from the target signal are important to the search process. In the kind of low signal-to-noise situation frequently encountered in the Shazam application, most of the tokens generated from the audio sample are garbage. But the presence of even a small percentage of good matching tokens is sufficient to flag a statistically significant probability of finding the correct song in a large database of songs.

We also found that the fingerprint features must be aligned temporally; that is, if a set of features appears in both the original recording in the database and in a sample query, the relative positions of each feature within each recording must be the same. Unlike speech recognition, in which dynamic time warping may be used to match up loosely corresponding features with a time-varying and nondeterministic rate of progression, the Shazam technique assumes the correspondence is directly linear; that is, if you plot the relative times of occurrence of each token in a time-versus-time scatterplot, a valid match should have points that accumulate on a straight diagonal line.

Such a line is quickly detected by searching for a peak in a histogram of relative time differences. This assumption concerning temporal alignment greatly accelerates the matching process and strengthens the acceptance/rejection criteria determining whether a given fingerprint feature is valid, thus providing a quick and effective way to filter out the large number

[1] An overlapping Short-Time Fourier Transform is calculated at regular intervals on the audio data, and a power level is calculated for each resulting time-frequency bin. A bin is a peak if its power level is greater than all the other bins in a bounded region around the bin.

TVision Insights, Inc. – Ex. 1026, Page 000004

THE PRESENCE OF corresponding spectrogram peaks in a noisy versus noiseless music signal makes it possible to determine with high probability whether an audio sample matches a recording in the database.

of garbage tokens generated in the combinatorial hashing stage.

We implemented the recognition algorithm in C++ with certain speed-critical sections optimized in assembly. The recognition server is implemented as a cluster of a few dozen commodity off-the-shelf 64-bit x86-based servers, each with 8GB of RAM and running an optimized Linux kernel.

As of June 2006, the Shazam database contained more than three million tracks. Each incoming request is received by a master process that then broadcasts the query to a farm of slave processors, each holding a piece of the database index in memory. Each slave independently searches its chunk of the universe of fingerprint tokens and reports its identification results to the master. The master collects the results and returns a report (concerning recognition) to the remote client. For a discussion on performance characteristics of the recognition algorithm, see [5].

The Shazam music recognition service (www.shazam.com) has been publicly available on mobile phones in the U.K. since 2002. Since then, it has expanded into more than 20 countries hosted through a variety of local partners under various service brands, including Verizon Wireless and Cingular in the U.S. As of June 2006, nearly six million paying customers worldwide had used the service.

Meanwhile, Philips Electronics demonstrated its own "robust hashing" audio fingerprinting algorithm in 2001. Like Shazam, it is capable of free-field audio identification (QBE), though it forms hashes from differential energy flux in a time-frequency grid [2]. This technology was acquired in 2005 by Gracenote, a music database company in Emeryville, CA. Also in 2001, the Fraunhofer Institut in Erlangen, Germany (www.iis.fraunhofer.de/), demonstrated a technique based on "spectral flatness" [3].

QBE music recognition is a commercial reality. In the next few years, as more carriers and phone manufacturers offer related services, QBE will likely become part of the standard mobile phone feature set, like camera phones are today. The cost per query should drop, and more integration into follow-on sales and discovery services should also be possible. Other search modalities (such as query-by-humming or -similarity) may also be added. C

References

1. Bond, P. StarCD: A star is born nationally seeking stellar CD sales. *Hollywood Reporter CCCLX,* 13 (Nov. 1, 1999), 3.
2. Haitsma, J., Kalker, T., and Oostveen, J. Robust audio hashing for content identification. In *Proceedings of the International Workshop on Content-based Multimedia Indexing* (Brescia, Italy, Sept. 19–21, 2001).
3. Herre, J., Allamanche, E., and Helmuth, O. Robust matching of audio signals using spectral flatness features. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (Mohonk, NY, 2001), 127–130.
4. Kenyon, S., Simkins, L., Brown, L., and Sebastian, R. *U.S. Patent 4,450,531: Broadcast Signal Recognition System and Method.* U.S. Patent and Trademark Office, Washington, D.C.; www.uspto.gov.
5. Wang, A. An industrial-strength audio search algorithm. In *Proceedings of the Fourth International Conference on Music Information Retrieval* (Baltimore, Oct. 26–30, 2003); www.ismir.net.
6. Wold, E., Blum, T., Keislar, D., and Wheaton, J. Content-based classification, search, and retrieval of audio. *IEEE Multimedia 3,* 3 (Fall 1996), 27–36.

Avery Wang (avery@shazamteam.com) is the chief scientist of Shazam Entertainment, Ltd., London, U.K.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2006 ACM 0001-0782/06/0800 $5.00

TVision Insights, Inc. – Ex. 1026, Page 000005