A Review of Algorithms for Perceptual Coding of Digital Audio Signals [†]

Ted Painter, *Student Member IEEE*, and Andreas Spanias, *Senior Member IEEE*

Department of Electrical Engineering, Telecommunications Research Center
Arizona State University, Tempe, Arizona 85287-7206
spanias@asu.edu, painter@asu.edu

ABSTRACT

During the last decade, CD-quality digital audio has essentially replaced analog audio. During this same period, new digital audio applications have emerged for network, wireless, and multimedia computing systems which face such constraints as reduced channel bandwidth, limited storage capacity, and low cost. These new applications have created a demand for high-quality digital audio delivery at low bit rates. In response to this need, considerable research has been devoted to the development of algorithms for perceptually transparent coding of high-fidelity (CD-quality) digital audio. As a result, many algorithms have been proposed, and several have now become international and/or commercial product standards. This paper reviews algorithms for perceptually transparent coding of CD-quality digital audio, including both research and standardization activities. The paper is organized as follows. First, psychoacoustic principles are described with the MPEG psychoacoustic signal analysis model 1 discussed in some detail. Then, we review methodologies which achieve perceptually transparent coding of FM- and CD-quality audio signals, including algorithms which manipulate transform components and subband signal decompositions. The discussion concentrates on architectures and applications of those techniques which utilize psychoacoustic models to exploit efficiently masking characteristics of the human receiver. Several algorithms which have become international and/or commercial standards are also presented, including the ISO/MPEG family and the Dolby AC-3 algorithms. The paper concludes with a brief discussion of future research directions.

I. INTRODUCTION

Audio coding or *audio compression* algorithms are used to obtain compact digital representations of high-fidelity (wideband) audio signals for the purpose of efficient transmission or storage. The central objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent signal reproduction, i.e., while generating output audio which cannot be distinguished from the original input, even by a sensitive listener ("golden ears"). This paper gives a review of algorithms for transparent coding of high-fidelity audio.

The introduction of the compact disk (CD) in the early eighties [1] brought to the fore all of the advantages of digital audio representation, including unprecedented high-fidelity, dynamic range, and robustness. These advantages, however, came at the expense of high data rates. Conventional CD and digital audio tape (DAT) systems are typically sampled at 44.1 or 48 kilohertz (kHz), using pulse code modulation (PCM) with a sixteen bit sample resolution. This results in uncompressed data rates of 705.6/768 kilobits per second (kbps) for a monaural channel, or 1.41/1.54 megabits per second (Mbps) for a stereo pair at 44.1/48 kHz, respectively. Although high, these data rates were accommodated successfully in first generation digital audio applications such as CD and DAT. Unfortunately, second generation multimedia applications and wireless systems in particular are often subject to bandwidth or cost constraints which are incompatible with high data rates. Because of the success enjoyed by the first generation, however, end users have come to expect "CD-quality" audio reproduction from any digital system. New network and wireless multimedia digital audio systems, therefore, must reduce data rates without compromising reproduction quality. These and other considerations have motivated considerable research during the last decade towards formulation of compression schemes which can satisfy simultaneously the conflicting demands of high compression ratios and transparent reproduction quality for high-fidelity audio signals [2][3][4][5][6][7][8][9][10][11]. As a result, several standards have been developed [12][13][14][15], particularly in the last five years [16][17][18][19], and several are now being deployed commercially [94] [97][100][102] (Table 2).

A. *GENERIC PERCEPTUAL AUDIO CODING ARCHITECTURE*

This review considers several classes of analysis-synthesis data compression algorithms, including those

[†] Portions of this work have been sponsored by a grant from the NDTC Committee of the Intel Corporation Direct all communications to A Spanias

which manipulate: transform components, time-domain sequences from critically sampled banks of bandpass filters, linear predictive coding (LPC) model parameters, or some hybrid parametric set. We note here that although the enormous capacity of new storage media such as Digital Versatile Disc (DVD) can accommodate *lossless* audio coding [20][21], the research interest and hence all of the algorithms we describe are *lossy* compression schemes which seek to exploit the psychoacoustic principles described in section two. Lossy schemes offer the advantage of lower bit rates (e.g., less than 1 bit per sample) relative to lossless schemes (e.g., 10 bits per sample). Naturally, there is a debate over the quality limitations associated with lossy compression. In fact, some experts believe that *uncompressed* digital CD-quality audio (44.1 kHz/16b) is intrinsically inferior to the analog original. They contend that sample rates above 55 kHz and word lengths greater than 20 bits [21] are necessary to achieve transparency in the absence of any compression. It is beyond the scope of this review to address this debate.

Before considering different classes of audio coding algorithms, it is first useful to note the architectural similarities which characterize most perceptual audio coders. The lossy compression systems described throughout the remainder of this review achieve coding gain by exploiting both *perceptual irrelevancies* and *statistical redundancies*. All of these algorithms are based on the generic architecture shown in Fig. 1. The coders typically segment input signals into quasi-stationary frames ranging from 2 to 50 milliseconds in duration. A time-frequency analysis section then decomposes each analysis frame. The time/frequency analysis approximates the temporal and spectral analysis properties of the human auditory system. It transforms input audio into a set of parameters which can be quantized and encoded according to a perceptual distortion metric. Depending on overall system objectives and design philosophy, the time-frequency analysis section might contain a

- Unitary transform
- Time-invariant bank of uniform bandpass filters
- Time-varying (signal-adaptive), critically sampled bank of non-uniform bandpass filters
- Hybrid transform/filterbank signal analyzer
- Harmonic/sinusoidal analyzer
- Source-system analysis (LPC/Multipulse excitation)

The choice of time-frequency analysis methodology always involves a fundamental tradeoff between time and frequency resolution requirements. Perceptual distortion control is achieved by a psychoacoustic signal analysis section which estimates signal masking power based on psychoacoustic principles (see section two). The psychoacoustic model delivers masking thresholds which quantify the maximum amount of distortion that can be injected at each point in the time-frequency plane during quantization and encoding of the time-frequency parameters without introducing audible artifacts in the reconstructed signal. The psychoacoustic model therefore allows the quantization and encoding section to exploit perceptual irrelevancies in the time-frequency parameter set. The quantization and encoding section can also exploit statistical redundancies through classical techniques such as differential pulse code modulation (DPCM) or adaptive DPCM (ADPCM). Quantization might be uniform or pdf-optimized (Lloyd-Max), and it might be performed on either scalar or vector quantities (VQ). Once a quantized compact parametric set has been formed, remaining redundancies are typically removed through run-length (RL) and entropy (e.g. Huffman, arithmetic, LZW) coding techniques. Since the psychoacoustic distortion control model is signal adaptive, most algorithms are inherently variable rate. Fixed channel rate requirements are usually satisfied through buffer feedback schemes, which often introduce encoding delays.

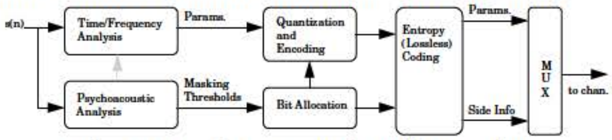


Fig. 1. Generic Perceptual Audio Encoder

The study of perceptual entropy (PE) suggests that transparent coding is possible in the neighborhood of 2 bits per sample [45] for most for high-fidelity audio sources (~88 kpbs given 44.1 kHz sampling). The lossy perceptual coding algorithms discussed in the remainder of this paper confirm this possibility. In fact, several coders approach transparency in the neighborhood of 1 bit per sample. Regardless of design details, all perceptual audio coders seek to achieve transparent quality at low bit rates with tractable complexity and manageable delay. The discussion of algorithms given in sections three through five brings to light many of the tradeoffs involved with the various coder design philosophies.

B. PAPER ORGANIZATION

The rest of the paper is organized as follows. In section II, psychoacoustic principles are described which can be exploited for significant coding gain. Johnston's notion of perceptual entropy is presented as a measure of the fundamental limit of transparent compression for audio. Sections III through V review state-of-the-art algorithms which achieve transparent coding of FM- and CD-quality audio signals, including several techniques which are established in international standards. Transform coding methodologies are described in section III, and subband coding algorithms are addressed in section IV. In addition to methods based on uniform bandwidth filterbanks, section IV covers coding methods which utilize discrete wavelet transforms and non-uniform filterbanks. Finally, section V is concerned with standardization activities in audio coding. It describes recently adopted standards including the ISO/IEC MPEG family, the Phillips' Digital Compact Cassette (DCC), the Sony Minidisk, and the Dolby AC-3 algorithms. The paper concludes with a brief discussion of future research directions.

For additional information, one can also refer to informative reviews of recent progress in wideband and hi-fidelity audio coding which have appeared in the literature. Discussions of audio signal characteristics and the application of psychoacoustic principles to audio coding can be found in [22],[23], and [24]. Jayant, *et al.* of Bell Labs also considered perceptual models and their applications to speech, video, and audio signal compression [25]. Noll describes current algorithms in [26] and [27], including the ISO/MPEG audio compression standard.

II. PSYCHOACOUSTIC PRINCIPLES

High precision engineering models for high-fidelity audio currently do not exist. Therefore, audio coding algorithms must rely upon generalized receiver models to optimize coding efficiency. In the case of audio, the receiver is ultimately the human ear and sound perception is affected by its masking properties. The field of psychoacoustics [28][29][30][31][32][33][34] has made significant progress toward characterizing human auditory perception and particularly the time-frequency analysis capabilities of the inner ear. Although applying perceptual rules to signal coding is not a new idea [35], most current audio coders achieve compression by exploiting the fact that "irrelevant" signal information is not detectable by even a well trained or sensitive listener. Irrelevant information is identified during signal analysis by incorporating into the coder several psychoacoustic principles, including absolute hearing thresholds, critical band frequency analysis, simultaneous masking, the spread of masking along the basilar membrane, and temporal masking. Combining these psychoacoustic notions with basic properties of signal quantization has also led to the development of perceptual entropy [36], a quantitative estimate of the fundamental limit of transparent audio signal compression. This section reviews psychoacoustic fundamentals and perceptual entropy, then gives as an application example some details of the ISO/MPEG psychoacoustic model one.

A. *ABSOLUTE THRESHOLD OF HEARING*

The absolute threshold of hearing is characterized by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The frequency dependence of this threshold was quantified as early as 1940, when Fletcher [28] reported test results for a range of listeners which were generated in an NIH study of typical American hearing acuity. The quiet threshold is well approximated [37] by the nonlinear function

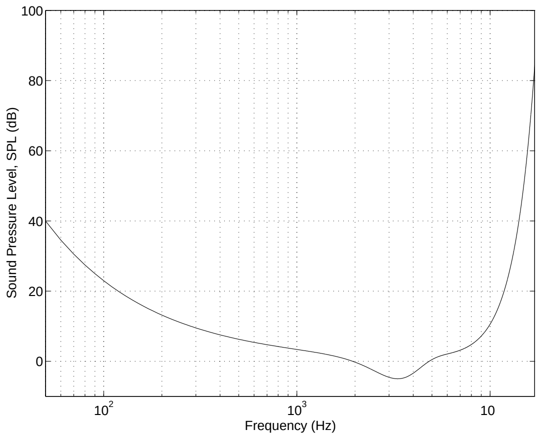


Fig. 2. The Absolute Threshold of Hearing

$$\begin{aligned}T_q(f) = 3.64(f/1000)^{-0\;8} \\ -6.5e^{-0\;6(f/1000-3\;3)^2} \quad \text{(dB SPL)} \qquad (1) \\ +10^{-3}(f/1000)^4\end{aligned}$$

which is representative of a young listener with acute hearing. When applied to signal compression, $T_q(f)$ can be interpreted as a maximum allowable energy level for coding distortions introduced in the frequency domain (Fig. 2). Algorithm designers have no *a priori* knowledge regarding actual playback levels, therefore the sound pressure level (SPL) curve is often referenced to the coding system by equating the lowest point on the curve (i.e., 4 kHz) to the energy in +/- 1 bit of signal amplitude. Such a practice is common in algorithms which utilize the absolute threshold of hearing.

B. *CRITICAL BANDS*

Using the absolute threshold of hearing to shape the coding distortion spectrum represents the first step towards perceptual coding. Next we consider how the ear actually does spectral analysis. It turns out that a frequency-to-place transformation takes place in the inner ear, along the basilar membrane. Distinct regions in the cochlea, each with a set of neural receptors, are "tuned" to different frequency bands. Empirical work by several observers led to the modern notion of critical bands [28][29][30][31] which correspond to these cochlear regions. In the experimental sense, critical bandwidth can be loosely defined as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise source at constant sound pressure level remains constant even as the bandwidth is increased up to the critical bandwidth. The loudness then begins to increase. In a different experiment (Fig 3a), the detection threshold for a narrowband noise source between two masking tones remains constant as long as the frequency separation between the tones remains within a critical bandwidth. Beyond this bandwidth, the threshold rapidly decreases (Fig 3c).

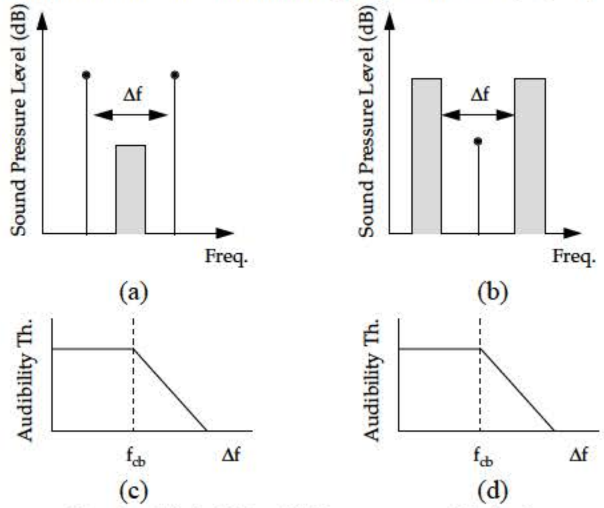
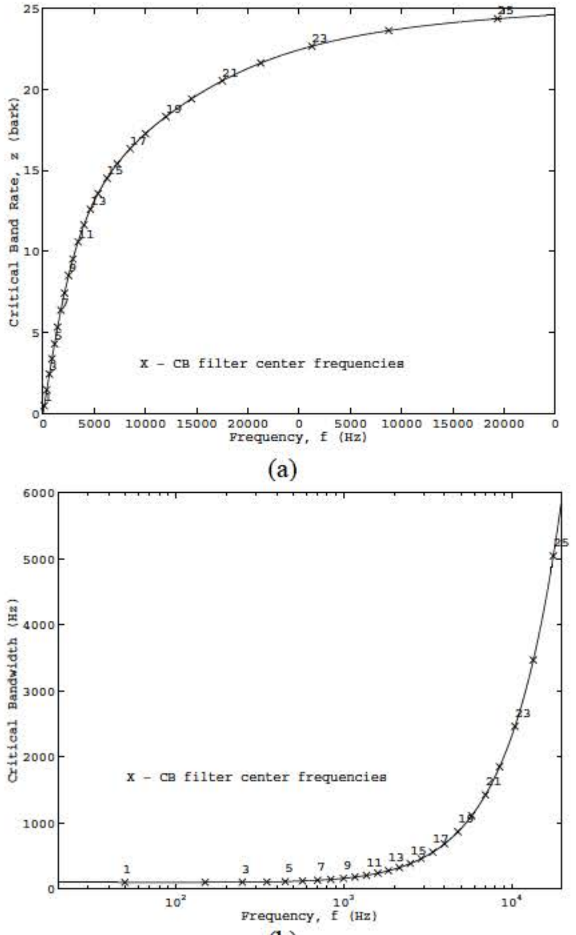


Fig. 3. Critical Band Measurement Methods



(a)

(b)

Fig. 4. (a) Critical Band Rate, *z(f)*, and (b) Critical Bandwidth, BW_c

A similar notched-noise experiment can be constructed with masker and maskee roles reversed (Fig. 3b,d). Critical bandwidth tends to remain constant (about 100 Hz) up to 500 Hz, and increases to approximately 20% of the center frequency above 500 Hz. For an average listener, critical bandwidth (Fig. 4b) is conveniently approximated [33] by

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \text{(Hz)} \quad (2)$$

Although the function BW_c is continuous, it is useful when building practical systems to treat the ear as a discrete set of bandpass filters which obeys Eq. (2). Table 1 gives an idealized filterbank which corresponds to the discrete points labeled on the curve in Figs. 4a, 4b. A distance of 1 critical band is commonly referred to as "one bark" in the literature. The function [33]

$$z(f) = 13\arctan(.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{(Bark)} \quad (3)$$

is often used to convert from frequency in Hertz to the bark scale (Fig 4a). Corresponding to the center frequencies of the Table 1 filterbank, the numbered points in Fig. 4a illustrate that the non-uniform Hertz spacing of the filterbank (Fig. 5) is actually uniform on a bark scale. Thus, one critical bandwidth comprises one bark. Intra-band and inter-band masking properties associated with the ear's critical band mechanisms are routinely used by modern audio coders to shape the coding distortion spectrum. These masking properties are described next.

Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700
18	4000	3700-4400
19	4800	4400-5300
20	5800	5300-6400
21	7000	6400-7700
22	8500	7700-9500
23	10,500	9500-12000
24	13,500	12000-15500
25	19,500	15500-

Table 1 Critical Band Filterbank [after Scharf]

C. SIMULTANEOUS MASKING AND THE SPREAD OF MASKING

Masking refers to a process where one sound is rendered inaudible because of the presence of another sound. Simultaneous masking refers to a frequency-domain phenomenon which has been observed within critical bands (in-band). For the purposes of shaping coding distortions it is convenient to distinguish between two types of simultaneous masking, namely *tone-masking-noise* [31], and *noise-masking-tone* [32]. In the first case, a tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape, provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone. The second masking type follows the same pattern with the roles of masker and maskee reversed. A simplified explanation of the mechanism underlying both masking phenomena is as follows. The presence of a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to effectively block transmission of a weaker signal. Inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. This effect, also known as the spread of masking, is often modeled in coding applications by an approximately triangular spreading function which has slopes of +25 and -10 dB per bark. A convenient analytical expression [35] is given by:

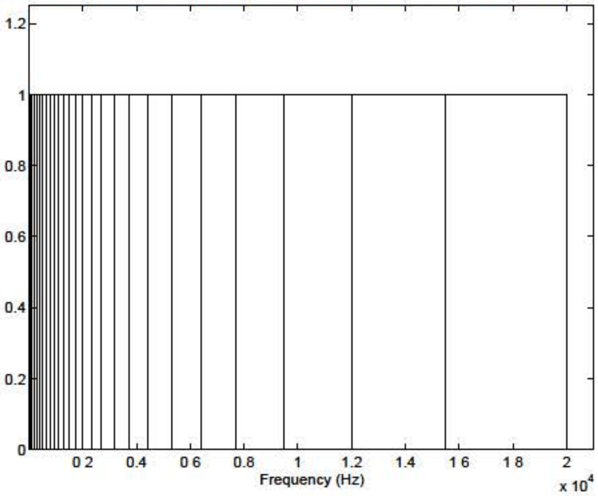


Fig. 5. Idealized Critical Band Filterbank

$$SF_{dB}(x) = 15.81 + 7\ 5(x+0.474) - 17\ 5\sqrt{1+(x+0.474)^2}\ \text{dB} \qquad (4)$$

where x has units of barks and $SF_{db}(x)$ is expressed in dB. After critical band analysis is done and the spread of masking has been accounted for, masking thresholds in psychoacoustic coders are often established by the [38] decibel (dB) relations:

$$TH_N = E_T - 14.5 - B \qquad (5)$$

$$TH_T = E_N - K \qquad (6)$$

where TH_N and TH_T, respectively, are the noise and tone masking thresholds due to tone-masking noise and noise-masking-tone, E_N and E_T are the critical band noise and tone masker energy levels, and B is the critical band number. Depending upon the algorithm, the parameter K has typically been set between 3 and 5 dB. Masking thresholds are commonly referred to in the literature as (bark scale) functions of just noticeable distortion (JND). One psychoacoustic coding scenario might involve first classifying masking signals as either noise or tone, next computing appropriate thresholds, then using this information to shape the noise spectrum beneath JND. Note that the absolute threshold (T_{ABS}) of hearing is also considered when shaping the noise spectra, and that MAX(JND, T_{ABS}) is most often used as the permissible distortion threshold. Notions of critical bandwidth and simultaneous masking in the audio coding context give rise to some convenient terminology illustrated in Fig. 6, where we consider the case of a single masking tone occurring at the center of a critical band. All levels in the figure are given in terms of dB SPL. A hypothetical masking tone occurs at some masking level. This generates an excitation along the basilar membrane which is modeled by a spreading function and a corresponding *masking threshold*. For the band under consideration, the *minimum masking threshold* denotes the spreading function in-band minimum. Assuming the masker is quantized using an m-bit uniform scalar quantizer, noise might be introduced at the level m. *Signal-to-mask ratio* (SMR) and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels, respectively.

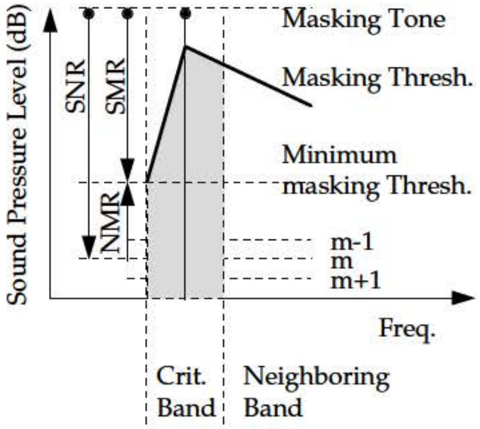


Fig. 6. Schematic Representation of Simultaneous Masking (after [26])

D. TEMPORAL MASKING

Masking also occurs in the time-domain. In the context of audio signal analysis, abrupt signal transients (e.g., the onset of a percussive musical instrument) create pre- and post- masking regions in time during which a listener will not perceive signals beneath the elevated audibility thresholds produced by a masker. The skirts on both regions are schematically represented in Fig. 7. In other words, absolute audibility thresholds for masked sounds are artificially increased prior to, during, and following the occurrence of a masking signal. Whereas premasking tends to last only about 5 ms,

DOCKET
ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.