Richard F. Lyon

Machine Hearing: An Emerging Field

If we had machines that could hear as humans do, we would expect them to be able to easily distinguish speech from music and background noises, to pull out the speech and music parts for special treatment, to know what direction sounds are coming from, to learn which noises are typical and which are noteworthy. Hearing machines should be able to organize what they hear; learn names for recognizable objects, actions, events, places, musical styles, instruments, and speakers; and retrieve sounds by reference to those names. These machines should be able to listen and react in real time, to take appropriate action on hearing noteworthy events, to participate in ongoing activities, whether in factories, in musical performances, or in phone conversations.

APPLICATIONS AND MOTIVATIONS

John Treichler's "Exploratory DSP" column "A View of the Future" [1] mentions a number of signal processing areas that are in the middle of a long trajectory of development; the sound-related ones include sonography, seismic exploration, telephony, music recording and compression, computer-laden automobiles, telepresence, speech synthesis and recognition, and sonar target detection and classification. Some of these (sonography, seismic, sonar) might be best served by techniques that have nothing to do with hearing. Others should benefit by an increased emphasis on hearing or on what things "sound like" to humans. Still other applications are not far enough along to make it on the list; for example, a very simple application that has been explored a bit in recent years is the personal audio diary: an audio recording of your daily life, which is now easy and inexpensive to capture and store, could be a great resource if there were good ways to analyze, organize, search, index, transcribe, and summarize it.

I envision a coming together of the telepresence, computer-laden car, speech, and music areas that Treichler mentions into a "smart environment" system that can converse with its occupants; keep track of things; serve as a security, surveillance, and diagnostic system; and provide entertainment and communication services. Since designing and building such a comprehensive system at this point is probably too big a job for anyone to take on, it might make sense to approach it instead by proliferating primitive hearing machines, which could be installed in cars, homes, meeting rooms, and portable computers, and open interfaces that would allow applications to be added incrementally to take advantage of these hearing front ends without reinventing or redeploying them. Obviously, such front ends would need to work well for speech, music, and all sorts of mixed environmental sounds, so a hearing-based approach is indicated.

Besides these real-time and interactive applications, there are lots of applications in the analysis of stored sound media. Our computers are presently mostly deaf, in that they have little idea what the sounds they store and serve represent. At Google, we store a lot of sound, including some speech databases, but mostly the unanalyzed sound tracks of videos. Wouldn't it make sense to have our computers listen to all of those and note what they're about, to categorize, organize, and index them? Not just what words are spoken, but what music is played, or what events and actions can be heard. The field of content-based analysis of images and videos has advanced steadily in recent years, but content-based analysis of sound tracks is somewhat lagging. Working video content-based analysis systems are low-hanging fruit for machine hearing, as sound features can easily make them better.

MACHINE HEARING AS A FIELD OF ENDEAVOR

Most reported work in sound analysis is applied to speech and music, but there is a much more general set of problems that are of interest here. We call this emerging, more general, field machine hearing. Compared to the diverse and active field of machine vision, the machine hearing field is still in its infancy, though the pieces of technology needed to move into diverse hearing applications are now mostly in hand. In this column, I discuss how I see this field developing, and how I see it addressing important current applications, and I make recommendations on strategies and approaches that I hope others will find useful to help advance the field.

In machine hearing, we focus on pragmatic system structures and real applications involving realistic sound mixtures in real environments. We hope to avoid the kind of split that the vision field had over the years, between "machine vision" in industry and "computer vision" in academia, and instead bring all the speech, music, and hearing researchers closer together by focusing on more general sound processing that provides a clear opportunity for leverage via collaboration.

In being pragmatic, we at the same time assume that machine hearing

Digital Object Identifier 10.1109/MSP.2010.937498

1053-5888/10/$26.00©2010IEEE

systems will work best when they hear like humans do, in the sense that they model the human hearing apparatus, part of which is shown in Figure 1, and that they create internal representations based on what things "sound like," as opposed to analyzing directly into representations of structures that make sound, such as vocal tracts. And we assume that the input sound will be a messy mixture, and so avoid representations that are optimized for one sound type or one sound source.

We hope and expect that machine hearing will emerge as a first-class academic and industrial field, much like machine vision and machine learning.

A MACHINE HEARING SYSTEM STRUCTURE

The machine hearing system structure that we are using as a baseline approach is one that we have modeled on some successful machine vision applications, and that has worked well for several sound-analysis applications already. Such a system consists of four main modules:

1) A peripheral analyzer: Common to all machine applications is a sound-analysis front end that models the amazing action of the cochlea in separating sounds into a set of overlapping bandpass channels, compressing the dynamic range of sounds, and producing a half-wave-rectified representation that preserves both the power and the fine time structure in all the channel waveforms.

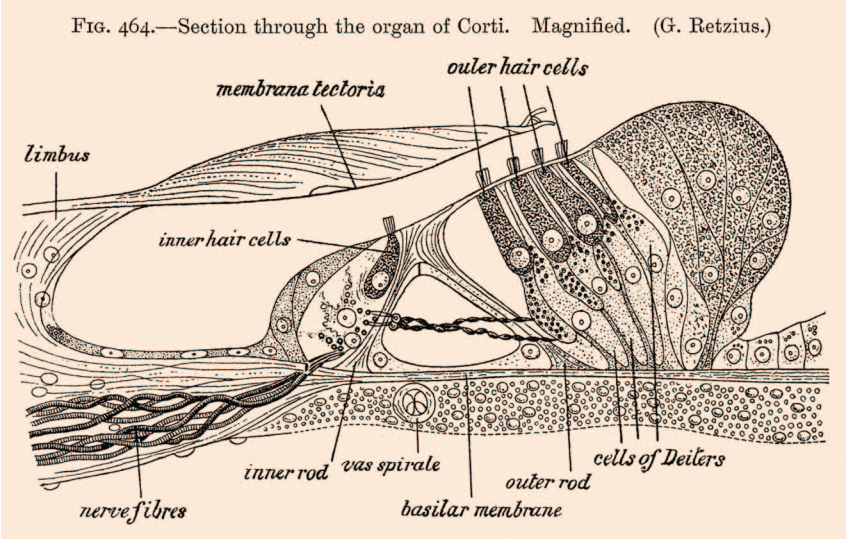
2) *One or more auditory image generators*: This stage demodulates fine temporal structure into more slowly changing representations, in the form of two-dimensional (2-D) moving image maps of the sort found in the auditory midbrain and projecting to the auditory cortex. For example, it generates a stabilized auditory image or correlogram, embodying joint spectral and temporal detail per Licklider's duplex theory of pitch perception [2], or a binaural correlogram per Jeffress's place theory of binaural localization [3].

3) A feature extraction module: As in machine vision systems, this stage gets moving (auditory) images as input and extracts the kinds of local and global (or multiscale) features that will work well with a following trainable classifier.

4) A trainable classifier or decision module: For the chosen application, appropriate machine learning techniques are applied to learn a mapping from the features extracted in the previous stages to the kinds of decisions needed by the application. This module can operate in a single step, as a single-layer perceptron does, or it can use or learn multiple layers of internal structure.

The first two modules respect human hearing, in the sense of having a goal of producing representations of what the sound stream "sounds like," while transforming the machine hearing problem into the form of a machine vision problem—reducing the machine hearing problem to the previously unsolved problem of machine vision, I've been told. Solved or not, this reduction allows useful leverage of successful techniques in the latter two modules, as well as plenty of room to improve at each stage.

Important concepts that can be shared between sound and image techniques include sparse representations, compression, multiscale analysis, three-dimensional (3-D) image-space motion analysis, and key-point detection, among others. For example, the representation might be sparsified as early as the output of the first module, where each half-wave waveform hump of bandpass filtered sound can be replaced by a discrete event indicating the time and size of the hump.



[FIG1] The organ of Corti, shown here as drawn by anatomist Gustaf Retzius circa 1880 and reproduced in *Gray's Anatomy*, is the inner ear's magical transducer assembly. The inner hair cells sense sound and drive most of the auditory nerve fibers, while the outer hair cells provide mechanical energy to amplify and compress traveling waves on the basilar membrane. The micromechanics of the organ of Corti is still an active area of research.

RESEARCH STRATEGY

The pursuit of auditory models for automatic speech recognition (ASR) has not been entirely successful, due to the highly evolved state of ASR system technologies, which are finely tuned to existing representations and to how phonetic properties of speech are manifest in those representations. Bourlard et al. [4] have made the point that if we're going to allow novel techniques such as auditory models into the ASR field, it will have to be done by tolerating a temporary performance setback and possibly slow recovery. A better strategy may be to look outside the mainstream ASR application, to applications that are so far under-served, and in which the typical speech models don't work well. To some extent, music applications may share similar drawbacks, since they tend to involve representations highly tuned to

the complex mathematical structures of musical pitch, rhythm, key and chord structures, etc. Therefore, we recommend the strategy of focusing on applications with mixed and unpredictable sound content, which can include some speech and music, but which are not competing directly with existing speech and music analysis systems.

Researchers need to address real problems, and evaluate and compare performance on such problems, with real-world noisy sound, to drive progress. Bake-offs such as the Music Information Retrieval Evaluation eXchange (MIREX) tasks are a great way to motivate conversion of ideas into running systems, and to get feedback on how they work, compared to what other researchers are doing. That approach helped the speech field advance, is helping the music field advance, and is needed to help more general machine hearing advance. Shared development and training data sets can be a useful part of this process, along with the competitive evaluations. It takes a community of some critical mass to have the will and the energy to organize such data sets and bake-offs, and that's something that I believe we're approaching, independent of the pure speech and music areas.

One particularly promising area of machine hearing research is computational auditory scene analysis (CASA). To the extent that we can analyze sound scenes into separate meaningful components, we can achieve an advantage in tasks involving processing of those components separately. Separating speech from interference is one such application. This concept has recently been applied by Audience Inc. to the problem of cleaning up the speech input to a mobile phone, in front of the voice coding [5]. Since the voice coders tend to work poorly on sound mixtures, but well on clean speech, there is good leverage here if interfering sounds can be suppressed at the input. The Audience technique uses a model of binaural hearing and treats the task as a CASA problem, with the result that the phone's coded speech sounds better and is more intelligible.

For many applications, however, a CASA approach is unnecessary even when the sound is a complicated mixture. Representations that give a good handle on what is in a mixture may be usable directly, without explicit identification of which parts of that mixture go together, or how many sources or streams are present. In the next section, we describe an example system that we implemented at Google, motivated by representations such as "bag of words" that have been useful in document analysis and retrieval, and the corresponding analogs that have been useful in image retrieval, even though documents and images contain arbitrary content mixtures.

The system structure that we described does not have a clear place for incorporating CASA, though the first two stages produce the sort of representations normally used in CASA, and the later stages do not preclude extracting stream-specific or source-specific features, or learning the properties of streams and sources. Strategically, we feel that CASA should remain on a research track for a few more years, while many applications can be addressed pragmatically without it in the short term.

With or without CASA, working with messy sound data is a strategic imperative, to force us to try to leverage what makes human hearing work so much better than systems that have been developed to work with clean speech and symbolic music.

Leveraging machine vision is another key strategy. Besides the use of ideas from the vision field, we can also leverage existing applications, as mentioned, by extending them to be audiovisual by simply adding sound features. We can do closer integrations, to leverage true audiovisual effects, for example in security and surveillance systems that include both cameras and microphones and need to track and identify what's going on. And we can extend ideas like visual tracking to more abstract kinds of sound tracking. Collaboration with machine vision researchers will help to grow the machine hearing field more quickly.

THE POLE–ZERO FILTER CASCADE PERIPHERAL MODEL

By analyzing a number of good properties that we want in model of the cochlea, or auditory periphery, we have converged on a pole–zero filter cascade (PZFC) structure as shown in Figure 2 [6]. This structure is based on fitting the magnitude and delay of basilar membrane traveling waves [7]. Recently, Mandal et al. have arrived at essentially the same cascaded pole–zero filter design by a more rigorous derivation based on models of impedance of the basilar membrane [8].

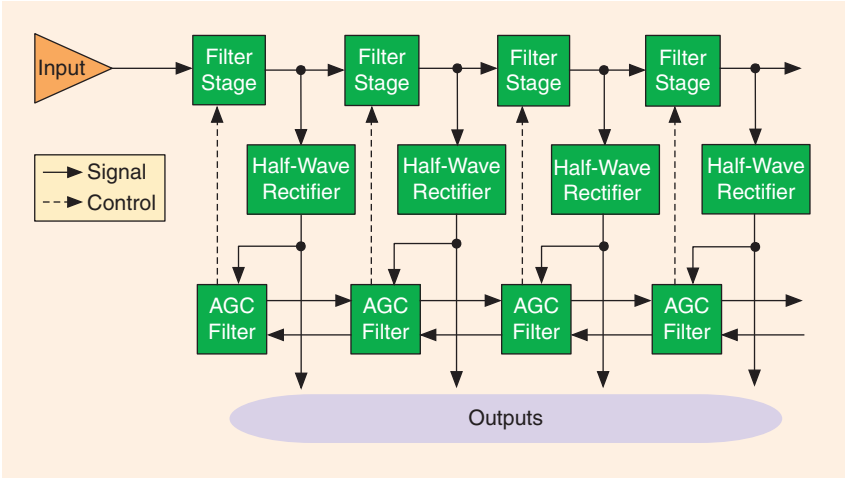
It is not necessarily important that the front-end filterbank be very true to the auditory system, but it probably can't hurt. By using a cascade structure that has a close connection to the underlying wave mechanics, and that provides excellent fits to both psychoacoustic and physiological data, we do get the strategic advantage of staying in closer connection with traditional hearing researchers who are advancing the understanding of the cochlea and other levels of the auditory system.

A key feature of the PZFC for machine hearing is its computational efficiency and simplicity, even while reproducing the complex nonlinear behavior of the magical transducer assembly of the organ of Corti. A cascade of simple second-order filter sections, one section per output channel, is nearly all there is to it. To get the nonlinearity, we add feedback control of parameters as a way to achieve an AGC for dynamic range compression. And we add an instantaneous cubic nonlinearity per stage, too, to give more very fast compression, and to generate realistic combination tones (Tartini's tones) that are known to be audibly propagating in the hydromechanics of the cochlea.

THE STABILIZED AUDITORY IMAGE

Our second module converts the signals on the auditory nerve to a more movie-like representation of the sort that is found in 2-D sheets of brain tissue as illustrated in Figure 3.

The stabilized auditory image is a representation developed by Patterson



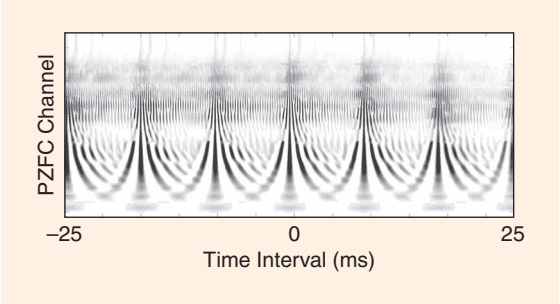
[FIG2] Schematic of the PZFC model of peripheral auditory filtering. The (top) cascaded filter stages provide a variable peak gain via a variable pole damping, which is adjusted by slowly varying feedback control signals from the (bottom) automatic gain control (AGC) smoothing network. The AGC loop corresponds to control to the cochlea's outer hair cell activity by efferent neurons from the olivary complex in the brainstem.

[9], closely related to the auditory correlogram [10], a realization of the duplex theory of pitch perception [2]. Patterson's image creation by triggered temporal integration is essentially a short-time cross-correlation of each channel's signal with a sparse trigger impulse signal, for trigger events chosen at prominent waveform peaks. He has experimented with a variety of modifications of the basic scheme, for example to create a scale-shift-covariant version designed to separate size effects from message effects in animal communications and human speech.

Many other imagelike or movielike auditory image representations are possible, for example to map interaural time difference and interaural level difference cues as computed in the brainsteam's olivary complex. In general, auditory images have an extra spatial dimension beyond the tonotopic or frequency dimension that's commonly used in various short-time spectral representations. This imagelike dimensionality is motivated by the 2-D structure of auditory cortex, and the various kinds of maps found in the auditory nervous system. As more is learned about these brain levels, we can expect to be able to incorporate the new knowledge into our machine hearing systems within the auditory image framework.

EXAMPLE SYSTEM: SOUND RETRIEVAL FROM TEXT QUERIES

Consider a large collection of sound files—potentially millions of sound effects, recordings, sound tracks, etc. It would be useful to find those files that are relevant to a user's text queries, such as "loud car crash." If we can learn a relationship between abstract sound features and query terms such as "loud," "car," and "crash" in a way that allows them to be naturally combined, we could support multiword queries effectively.



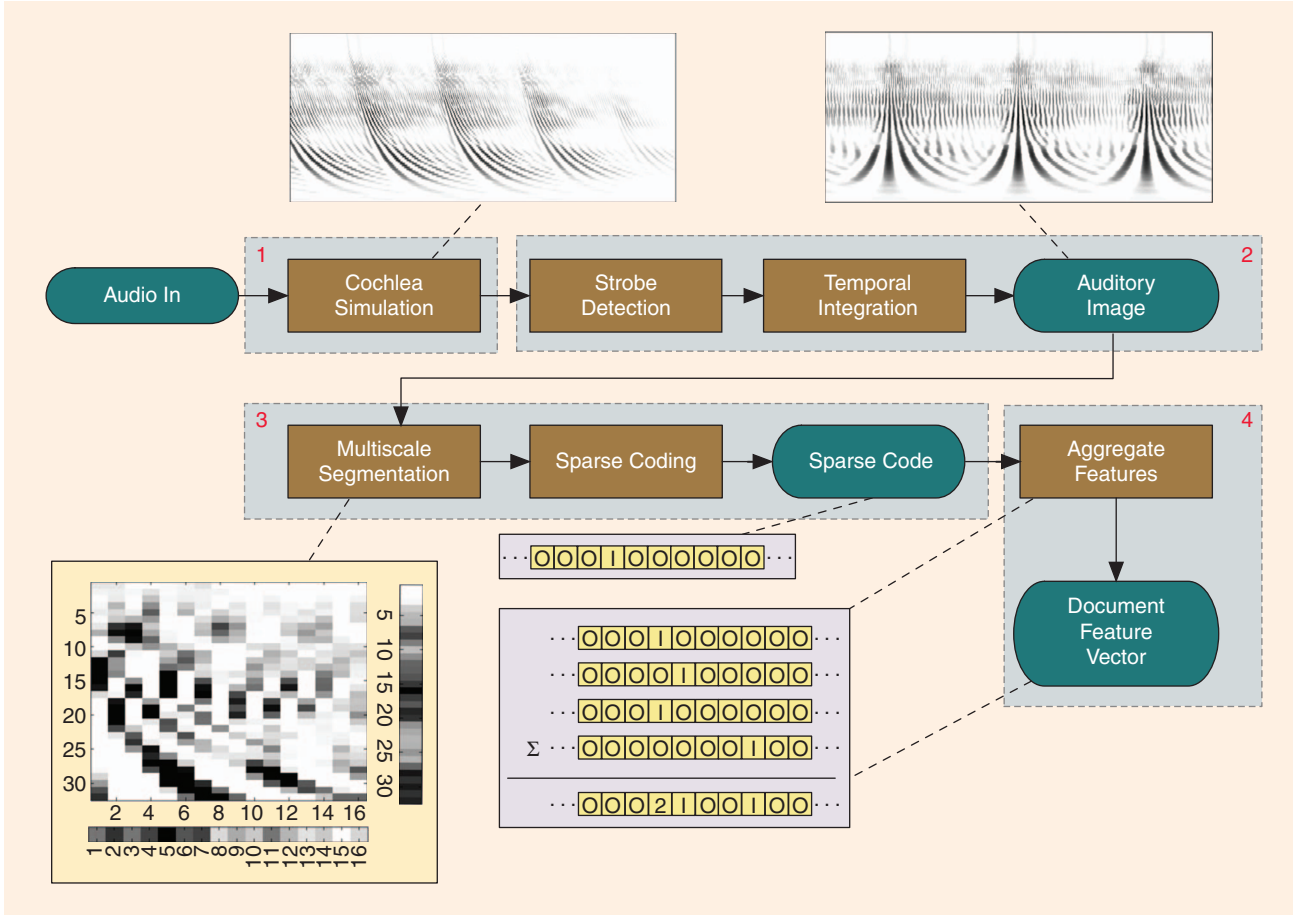
[FIG3] Example of an auditory image frame in response to a spoken vowel sound, using a very simple trigger detection method. The periodicity along the time lag dimension is a prominent feature of voiced speech, while the message, the vowel identity, is in the formants, the frequency bands in which the energy is concentrated. The image shows a low first formant, and high second and upper formants, indicating a high front vowel such as "ee."

This was the system that we selected for our first machine hearing experiments at Google, partly because of the availability of Grangier and Bengio's passive-aggressive method for image retrieval (PAMIR) technology that had been recently developed to do the same kind of thing for image retrieval [11].

The PAMIR method requires that our Stage 3 deliver a "bag of features" to represent a document (sound file, image file, or text file). A *bag* is like a *set* but with counts of repeated elements; equivalently, it is a histogram of how many times each feature occurs in the document. With sparse features, the bag is itself sparse; that is, since most counts are zero, only the nonzero counts need to be represented.

The abstract sparse features that worked well for image retrieval were multiscale abstract codes for local structure at locations all over the image. We made an analogous representation of the frames of an auditory image movie by using vector quantization (VQ), as shown in Figure 4, of many image patches of different sizes and aspect ratios. For example, a typical experiment configuration used 49 different patches, or boxes as we called them, each quantized through its own VQ codebook of 256 typical patterns specific to that box size and location, for a total of about 12,500 feature dimensions. At each frame, at a rate of 50 frames per second, only 1/256 of the features would be present; counting these occurrences over all the frames in a sound file usually resulted in 90% or more of the feature dimensions still being zero, so sparse representations were effective.

PAMIR uses a fast and robust training procedure to optimize a simple linear mapping from features to query terms, given training data with known tags. The query is represented as a sparse vector of terms in the tag vocabulary (about 3,000 words), and each sound file is given a score respect to a query via a linear matrix



[FIG4] Generating sparse codes from an "audio document," in four steps: 1) cochlea simulation, 2) stabilized auditory image creation, 3) sparse coding by vector quantization of multiscale patches, and 4) aggregation into a "bag of features" representation of the entire audio document. Steps 3 and 4 here correspond to the feature extraction module in the four-module system structure. To the fourth module, a PAMIR-based learning and retrieval system, this entire diagram represents a front end providing abstract sparse features for audio document characterization.

product of features times matrix times query. The matrix is trained to optimize a ranking criterion, such that it attempts to rank "relevant" documents higher, by giving them a higher score, than "non-relevant" ones, in the training set, for a large number of training queries that include multiword queries formed from the tag vocabulary.

The attractiveness of this approach was that we could use PAMIR for our Stage 4, since it didn't contain anything specific to images, and we could use a simple abstract VQ-based feature extraction for Stage 3, not tied to any particular sound classes or ideas of where in the auditory image the important distinguishing information might be. We compared the PAMIR approach to other trainable classifiers, support vector machines and mixture of Gaussians, and to another front-end representation vector quanitized mel-frequency cepstral coefficients (MFCCs). They all worked fairly well, but the PAMIR technique was much faster to train, and the auditory image features gave the best performance, if we increased the dimensionality by going to larger codebooks [12].

We are presently doing experiments with more challenging, but still controlled, sound mixtures for which we have known text tags, constructed for example by adding pairs of sound files together, and finding that the auditory sparse-coding approach shows an advantage in interference.

LEVERAGING MACHINE VISION AND MACHINE LEARNING

We have dozens of books with "machine vision" in the title, exploring techniques and applications. Each one can provide ideas and inspiration for machine hearing techniques and applications. Most applications are trainable, based on "machine learning." The game is mostly about how to extract features, from images or sounds, that work well with machine learning systems, and then train these systems to meet the needs of an application.

Some learning systems work best with fairly low feature dimensionality. ASR systems typically use a 39-dimensional MFCC-based feature vector, and learn distributions in feature space as mixtures of Gaussians. Other techniques, such as PAMIR from the vision field, deal best with very high feature dimensionality and don't try to model the distribution in feature space. By paying attention to what

(continued on page 139)

worst case complexity is the same as that of the exhaustive search, practically it is much faster than the exhaustive search. The binary code of the branch-and-bound search is downloadable from the Web page.

VOLUMETRIC FEATURES FOR EVENT DETECTION

http://www.cs.cmu.edu/~yke/video/

This method explores the use of volumetric features for event detection. It correlates spatiotemporal shapes to video clips that have been automatically segmented. As it works on over-segmented videos, background subtraction for reliable object segmentation is not required. A flow-based correlation technique is applied for matching, and can detect a wide range of actions in video. It can well handle the cluttered background. However, the detection speed is slow due to the large search space. This approach has relatively limited ability to handle action variations because only one action template is used. The code for feature extraction can be downloaded from the Web site.

SPACE-TIME SHAPE MATCHING FOR ACTION DETECTION

http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

Human action in video sequences can be seen as silhouettes of a moving torso and protruding limbs undergoing articulated motion. This method regards human actions as three-dimensional shapes induced by the silhouettes in the space-time volume. It is a generalization of a two-dimensional shape-analysis technique to the case of volumetric space-time shapes. The technique exploits Poisson equation solutions to extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation.

SUCCESSIVE CONVEX MATCHING FOR ACTION DETECTION

http://cs.bc.edu/~hjiang/scm/demo.html

In this approach, human actions are represented as sequences of postures. Each posture is represented as a transformed edge map. Specific actions are detected in a video by matching the time-coupled posture sequences to video frames. The template sequence to video registration is formulated as an optimal matching problem. A successive convex matching scheme is used to improve the matching speed. The demo code for successive convex matching can be accessed from the Web site.

AUTHORS

Junsong Yuan (jsyuan@ntu.edu.sg) is an assistant professor with the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore.

Zicheng Liu (zliu@microsoft.com) is a researcher with Microsoft Research, Redmond, Washington.

[SP]

[exploratory DSP] continued from page 135

techniques are working well in machine vision applications, we expect to continue to find good inspiration for what might work well for auditory-image-based machine hearing applications. When we find ideas worth trying, it may be easy to obtain implementations that can be adapted to use the output of our auditory analysis stages. Such repurposing of machine vision systems may provide good leverage in machine hearing research.

CONCLUSION

The machine hearing field is starting to find its feet. Applications are abundant and many are easy to address with known auditory front ends, combined with known feature extraction and machine learning techniques such as those that have proven successful in analogous applications in machine vision.

The signal processing technology involved is diverse but not too complex. Nonlinear filters, correlators, vector quantizers, and online learning algorithms, are involved in ways that can be initially fairly simple, yet leave room for open-ended research and improvement. Cooperation with researchers in auditory psychology and physiology will be highly valued on both ends.

Curing our machines' deafness, leveraging our knowledge of the amazing capabilities of the mammalian cochlea and auditory brain is a goal that will keep this field busy for a while and that will provide rewards on many fronts.

AUTHOR

Richard F. Lyon (dicklyon@ieee.org) is a research scientist at Google, Inc., and a Fellow of the IEEE.

REFERENCES

[1] J. Treichler, "Signal processing: A view of the future, Part I," *IEEE Signal Processing Mag.,* vol. 26, no. 2, pp. 116–120, 2009.

[2] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–133, 1951. Reprinted in *Physiological Acoustics,* E. D. Schubert, Ed. Stroudsburg, PA: Dowden, Hutchinson and Ross, Inc., 1979.

[3] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.,* vol. 41, no. 1, pp. 35–39, 1948.

[4] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Commun.,* vol. 18, no. 3, pp. 205–231, 1996.

[5] L. Watts, "Commercializing auditory neuroscience," in *Proc. Frontiers of Engineering: Reports on Leading-Edge Engineering 2006 Symp.*, 2007, p. 5.

[6] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proc. IEEE Int. Conf. Circuits and Systems*, 2010, pp. 3809–3812.

[7] R. F. Lyon, "Filter cascades as analogs of the cochlea," in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, T. S. Lande, Ed. Norwell, MA: Kluwer, 1998, pp. 3–18.

[8] S. Mandal, S. M. Zhak, and R. Sarpeshkar, "A bio-inspired active radio-frequency silicon cochlea," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, 2009, pp. 1814–1828.

[9] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Proc. 9th Int. Symp. Hearing, Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds. Oxford: Pergamon, 1992, pp. 429–446.

[10] M. Slaney and R. F. Lyon, "On the importance of time—A temporal representation of time," in *Visual Representations of Speech Signals,* M. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993, pp. 95–116.

[11] D. Grangier and S. Bengio, "A neural network to retrieve images from text queries," in *Proc. Artificial Neural Networks—ICANN 2006,* 2006, pp. 24–34.

[12] M. Rehn, R. F. Lyon, S. Bengio, T. C. Walters, and G. Chechik, "Sound ranking using auditory sparse-code representations," in *ICML Workshop Sparse Methods for Music Audio,* 2009.

[SP]