Using a Cascade of Asymmetric Resonators with Fast-Acting Compression as a Cochlear Model for Machine-Hearing Applications [*]

◯ Richard F. Lyon (Google Inc.)

Abstract

Every day, machines process many thousands of hours of audio signals through a realistic cochlear model. They extract features, inform classifiers and recommenders, and identify copyrighted material. The machine-hearing approach to such tasks has taken root in recent years, because hearing-based approaches perform better than we can do with more conventional sound-analysis approaches. We use a bio-mimetic "cascade of asymmetric resonators with fast-acting compression" (CAR-FAC)—an efficient sound analyzer that incorporates the hearing research community's findings on nonlinear auditory filter models and cochlear wave mechanics. The CAR-FAC is based on a pole–zero filter cascade (PZFC) model of auditory filtering, in combination with a multi-time-scale coupled automatic-gain-control (AGC) network. It uses simple nonlinear extensions of conventional digital filter stages, and runs fast due to its low complexity. The PZFC plus AGC network, the CAR-FAC, mimics features of auditory physiology, such as masking, compressive traveling-wave response, and the stability of zero-crossing times with signal level. Its output "neural activity pattern" is converted to a "stabilized auditory image" to capture pitch, melody, and other temporal and spectral features of the sound.

1 Introduction

Large-scale commercial applications of machine hearing are no longer just about speech. Sound similarity measures of various kinds, based on bio-mimetic auditory models, are used in applications such as music recommendation, content identification, and categorization of audio content. These applications are often integrated with similar video applications; for example, in analyzing YouTube videos and soundtracks.

To support machine-hearing applications, we have developed models of hearing that both run fast and realistically mimic the human cochlea. The *filter cascade* approach that we have developed is an efficient alternative to the more conventional parallel filterbank, and is also very closely connected to the way sound information propagates as traveling waves in the cochlea. We find that it also works well in sound-processing applications.

At the output of the cochlear model, information about sound is still encoded in the time domain, but spread across many frequency channels. The fine temporal structure of sound is extracted by further processing, mimicking the auditory brainstem, into movie-like representations known as stabilized auditory images, or correllograms. The frames of this movie are pictures that represent what a signal "sounds like." From these image frames, we extract the features that relate to tasks. Various kinds of machine learning then map these features to decisions, or to embeddings in a space with simple distance properties, to complete the task.

2 Modeling Cochlear Function

The PZFC is related to traveling-wave propagation in the cochlea, via methods inspired by the WKB approximation used in solving nonuniform distributed wave systems [1]. In the cascade of filters, each filter stage models a segment of the nonuniform distributed system. The stage transfer function is a pole–zero approximation to the transfer function corresponding to the local complex wavenumber. The cascade produces samples of the traveling wave at a discrete set of outputs, as shown in Fig. 1. Feedback from level detectors controls the tuning of the PZFC stages in response to sound level and spectrum, causing fast-acting compression. We refer to this combined filtering and compression model as the CAR-FAC.

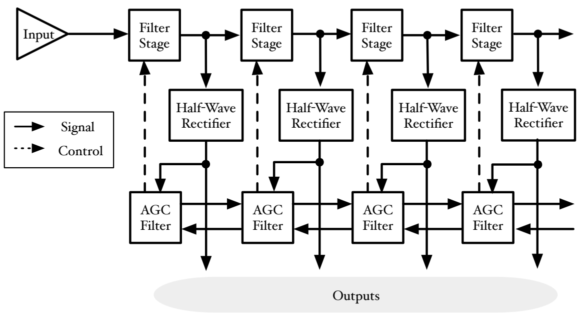
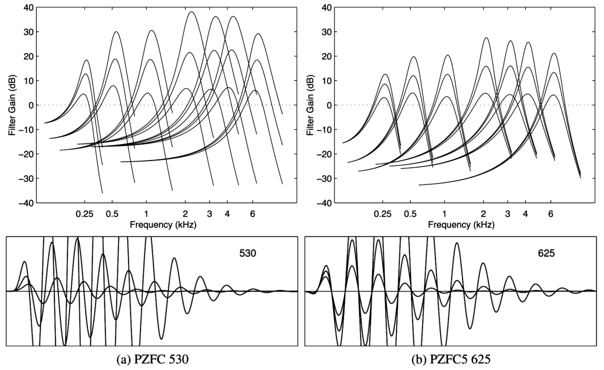


Fig. 1 Schematic of the CAR-FAC model of peripheral auditory filtering. The cascaded filter stages, or PZFC, (top) provide a variable peak gain via a variable pole damping. The pole damping is adjusted by slowly varying feedback control signals from the automatic gain control (AGC) smoothing network (bottom). The AGC loop corresponds to control of the cochlea's outer hair cell activity by efferent neurons from the olivary complex in the brainstem. Instantaneous local compression can also be included in the stages, to model the saturating activity of outer hair cells in the cochlea.

[*]



(a) PZFC 530

(b) PZFC5 625

Fig. 2 The transfer functions and impulse responses at one output of the PZFC, with the fixed zeros (left) and with movable zeros (right), are plotted for three different sound levels (levels corresponding to 30, 50, and 70 dB SPL tone detection threshold in broadband noise). The main differences are the high-side rolloff and the zero-crossing stability.

We have shown that the parameters of the model can be adjusted to fit human psychophysical data, in a task designed to show the properties of auditory filters over a range of frequencies and levels. Following Patterson, Irino, and Unoki [2, 3, 4], we use a nonlinear optimization procedure to fit the data from two labs [5, 6], covering many subjects and conditions, on the task of detecting tones in notched-spectrum masking noise. The result is that the PZFC fits the data better, with fewer parameters, than any previously considered models; although it fits best when the zeros are fixed, it still fits the data at least as well as previous models when the zeros are allowed to move, which is the condition needed to also match impulse-response or "revcor" data from the auditory nerve [7].

The model-parameter fitting procedure leads to detailed transfer functions and impulse responses, such as those illustrated in Fig. 2 for two different variants of the PZFC. The first version varies with level according to the pole motion shown in Fig. 3, along with the corresponding transfer function of a single stage from the cascade. Earlier and later stages look the same, but shifted to higher and lower frequencies, respectively. The other variant that we consider moves the zeros along with the poles, as shown in Fig. 4, corresponding to the right side of Fig. 2.

With the dynamic AGC connected, the dampings of the stages can vary somewhat independently, reducing the gain in those parts of the spectrum that have a lot of energy. The resulting shift in the transfer functions to a set of outputs is illustrated in Fig. 5, for the case of the model adapting to a speech sound.

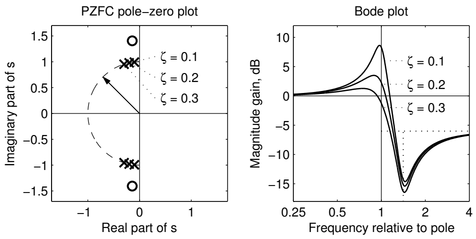


Fig. 3 Diagram of the motion of the poles of a PZFC stage in response to a gain-control feedback signal, and the effect on the resonator gain. The positions indicated by crosses at fixed radius (natural frequency) in the s-plane plot (left) correspond to pole damping ratios (ζ) of 0.1, 0.2, and 0.3, while the zero's damping ratio remains fixed at 0.1. Corresponding transfer function gains (right) of this asymmetric resonator stage do not change at low frequencies, but vary by several decibels near the pole frequency. The fact that the stage gain comes back up after the dip has little effect in the transfer function of a long cascade of such stages.

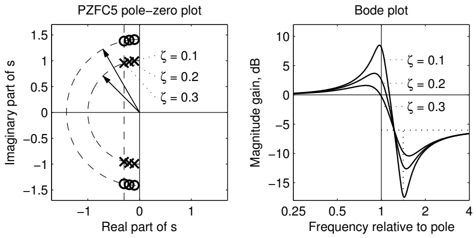


Fig. 4 The PZFC5 variant of the stage has the zeros moving along with the poles, giving it a slightly different high-frequency rolloff behavior: when the peak is high, the dip is deep. This is the version that more accurately reproduces observed stable zero-crossing behavior.

3 The Auditory Brain

Connecting the ear to high-level decision making, or extraction of meaning, requires at least another level or two of feature extraction. These layers can be considered to be abstract models of processing in the auditory brainstem and auditory cortex.

At the brainstem level, we assume the existence of structures that perform as hypothesized by Licklider in his "duplex theory of pitch perception", and as developed by Patterson as the "stabilized auditory image" (SAI). This mechanism converts the neural activity pattern from the peripheral model into a movie-like sequence of auditory images. See Fig. 6 for an example of one frame of an SAI. These images can be imagined as projecting to auditory cortex, in much the same way as visual images project from retina to visual cortex.

At the next level up, to abstractly model the cortex, we have adopted techniques used in machine vision, extracting local multi-scale sparse features via vector quantization of patterns in different re-

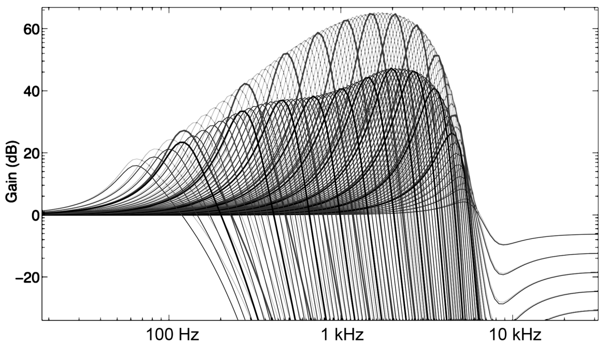


Fig. 5 CAR-FAC transfer functions when adapted to silence (higher curves) have very high peak gains, especially in the mid frequencies, due to the low pole dampings. When adapted to a moderate-level speech sound (lower curves), the gains reduce, compressing the dynamic range of the output compared to the input.

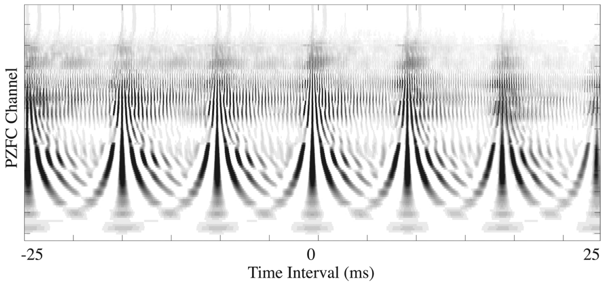


Fig. 6 Example of an auditory image frame in response to a spoken vowel sound. The periodicity along the time-lag dimension is a prominent feature of voiced speech, while the message, the vowel identity, is in the formants, the frequency bands in which the energy is concentrated. The image shows a low first formant, and high second and upper formants, indicating a high front vowel such as "ee."

gions of the SAI. The one-of-N code vectors from the VQ codebooks for each region are concatenated to make very large sparse-code vectors. These are accumulated over an entire sound file to make a "bag of features" representing the file, as summarized in Fig. 7. Using this feature vector, various machine-learning techniques are applied to make classifiers, application-specific distance measures or embeddings, and such.

4 Why It Works

The described models do well at many things. We understand some of these, such as why the CAR-FAC runs fast—the computational load is not much more than a second-order filter per output channel. And we believe that the CAR-FAC fits psychophysical data and neural revcor data well because it is closely related to wave propagation in the cochlea. But why do these models do well on machine-hearing tasks?

We have assumed that incorporating mechanisms and effects from biological hearing into our machine models will be helpful. Let us consider some of these mechanisms, and review why they might help.

The filter-cascade approach leads to realistic asymmetric transfer functions, and provides an easy way to incorporate dynamic level dependence as well as instantaneous distortion nonlinearities. The AGC filter network models the adaptation of cochlear gain, at several different time scales. In combination, these mechanisms do a good job of modeling at least some masking effects; the output of the model will therefore represent audible differences, and suppress inaudible differences. The coupled AGC tends to make the model response emphasize changes relative to recent spectral history, adapting out channel effects such as spectral tilt and overall loudness—as in "relative spectral" (RASTA) filtering of short-time spectral representations [8].

But the output of the cochlear model contains much more than spectral information. As Patterson et al. have shown, the SAI captures fine temporal structure in a way that explains many pitch phenomena, and in a way that seems optimized for recognizing and characterizing animal communication sounds, including speech and music [9, 10].

The extraction of features from the SAI is only very abstractly a bio-mimetic model, but can be viewed as analogous to the multiscale visual features thought to be extracted in early visual cortical areas. By using regions of different sizes and positions, we get features at different scales, and in different parts of the frequency/periodicity plane, such that sounds that are interfered with in one region may still come through by dominating in another region. Our reported experiments on sound retrieval in interference suggest that this approach pays off [11].

The high-dimensional sparse feature space that we construct by concatenating and counting sparse local codes allows linear decision boundaries to separate classes of sounds, in many cases. Thus, regularized linear machine-learning techniques can be applied efficiently and effectively. Even with feature spaces of 100,000 dimensions, and even with learning of 300 million coefficients to map abstract features to a vocabulary of 3000 words, training is fast and over-fitting is not a problem, if appropriate learning algorithms are used [12].

Besides the retrieval task, we have experimented with audio fingerprinting, cover-song detection, music recommendation, beat tracking, speech recognition, and a variety of classifiers, using the described features and other more specialized features. We have deployed several successful applications based on these experiments. Many of these tasks have the property that they don't need to work perfectly to be valuable, and they are more valuable as we make them work better. In many cases, making them work better seems to depend on representations of what the sounds "sound like," so we hope

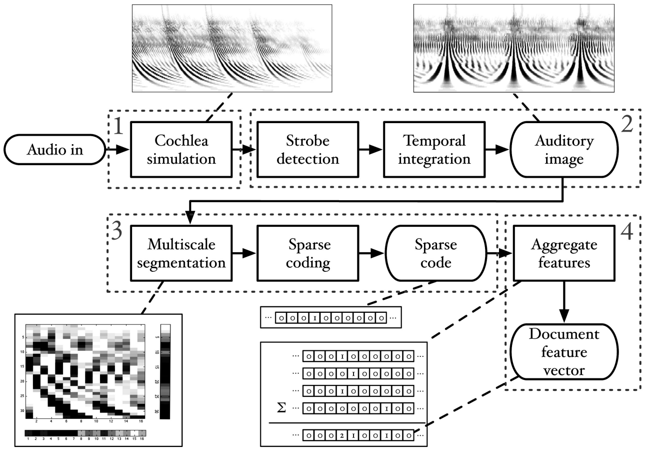


Fig. 7 Generating sparse features from an audio file, in four steps: (1) cochlea simulation using the CAR-FAC; (2) stabilized auditory image creation via triggered temporal integration; (3) sparse coding by vector quantization of multi-scale patches; (4) aggregation into a "bag of features" representation of the entire audio file. The application of this system to audio-file retrieval from text queries resulted in better performance than with other representations tried [12]. It was especially better for retrieving sounds mixed with other interfering sounds [11].

to find continuing success in using auditory models for these purposes.

5 Conclusion

The filter-cascade approach to cochlear modeling is efficient and realistic. It connects to the underlying mechanics, to psychophysical results, and to auditory-nerve physiology. Used at large scale to analyze, classify, and recognize sounds, it results in better accuracy than other sound feature-extraction approaches that we have tried. We are working to determine precisely what factors contribute to its success. Results so far are encouraging for this biomimetic approach to machine hearing, and machine perception more generally. More work to understand and improve this approach is needed, and we invite others to join in and help.

Acknowledgment Special thanks are due to Irino, Unoki, and Patterson for leading the way on the design and parameter fitting of auditory filter models, and on the design of sound-processing algorithms based on models of hearing; and to Baker, Rosen, Darling, Glasberg, and Moore for sharing the experimental data used in these efforts, and now in mine; and to my colleagues at Google—Chechik, Rehn, Bengio, Walters, Ponte, and others—for all their help in the experiments and applications.

References

[1] R. F. Lyon, "Filter cascades as analogs of the cochlea", in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, edited by T. S. Lande, 3–18 (Kluwer Academic Publishers, Norwell, Mass.) (1998).

[2] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp", J. Acoust. Soc. Am. **101**, 412–419 (1997).

[3] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter", J. Acoust. Soc. Am. **114**, 1529–1542 (2003).

[4] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore, and R. D. Patterson, "Comparison of the roex and gammachirp filters as representations of the auditory filter", J. Acoust. Soc. Am. **120**, 1474–1492 (2006).

[5] R. J. Baker, S. Rosen, and A. M. Darling, "An efficient characterisation of human auditory filtering across level and frequency that is also physiologically reasonable", in *Psychophysical and Physiological Adv Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, 81–88 (Whurr) (1998).

[6] B. R. Glasberg and B. C. J. Moore, "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise", J. Acoust. Soc. Am. **108**, 2318–2328 (2000).

[7] R. F. Lyon, "A pole–zero filter cascade provides good fits to human masking data and to basilar membrane and neural data", in *Mechanics of Hearing, 2011 Workshop*, edited by C. Shera et al. (AIP) (in press).

[8] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech and Audio Proc. **2**, 578–589 (1994).

[9] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images", in *Auditory physiology and perception, Proceedings of the 9th International Symposium on Hearing*, edited by Y. Cazals, L. Demany, and K. Horner, 429–446 (Pergamon, Oxford) (1992).

[10] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images", Advances in Speech, Hearing and Language Processing **3**, 547–563 (1996).

[11] R. F. Lyon, J. Ponte, and G. Chechik, "Sparse coding of auditory features for machine hearing in interference", in *IEEE Intl. Conf. Acoustics Speech and Signal Proc.*, 5876–5879 (2011).

[12] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, "Sound retrieval and ranking using sparse auditory representations", Neural computation **22**, 2390–2416 (2010).