Auditory Effects for ASR

by Richard F. Lyon -- Apple Computer, Inc. -- lyon@apple.com

13 June 1994 -- revised 23 April 1996 for ICASSP96

I briefly (and informally) summarize some of the kinds of auditory or psychoacoustic effects that one might want to try to exploit in automatic speech recognition (ASR). It's an admittedly Lyon-centric view, attepting to justify all this auditory modeling stuff I do in terms of its potential impact on ASR.

General rule: strong sounds mask weak sounds, "capturing" internal represenations and perception, when the properties of the sounds are otherwise similar; but weaker sounds can break through and be perceived if they are separated from the strong sounds along one or more important dimensions (time, frequency, pitch, direction, etc.).

ASR motivation: modeling masking can help reduce the sensitivity of internal representations to small noises, without reducing the sensitivity to subtle but perceptible and important speech cues. (Auditory masking models are now used in high-quality audio coders, such as MPEG-Audio).

Note: these effects are described very informally, with the notion of "salient" being used imprecisely, intending to indicate either what part of a signal's features are robustly represented or are easily accessible to higher levels. It would be a bigger job to track down and document exactly what effects have been measured for each of these items.

Items 1-6 have simple implications for a spectral-slice style ASR front end.

Items 7-11 define the potential for a longer-term auditory approach to signal separation, grouping, enhancement, streaming, selective attention, or whatever you want to call it.

For each numbered item, we can look at how the effect relates to typical ASR front-end representations, and how an auditory approach might lead to an improvement:

1. Perceived sound quality is almost independent of loudness, if loud enough.

Almost all ASR front ends use an amplitude-independent representation of spectral shape as the primary feature vector, obtained via some combination of normalization, logarithms, or AR modeling. They also typically represent total power or loudness as a separate feature. These ideas are fine to first order, and have gotten ASR to where it is today. But they totally punt on the issue of what is "loud enough".

2. Short relatively silent intervals are perceived as such, independent of the spectrum of the low-level signal in that interval.

This is the flip side of item 1. Intervals that are not "loud enough" to lead to a robust

representation of spectral shape are on their way toward true silence. Perceptually, all silences are alike, independent of the sprectrum of the residual noise, if they are "weak enough" relative to nearby loud parts. A representation that approaches the conventional normalized form for loud signals but gracefully denormalizes toward a stable representation of silence would be an improvement. But then somehow the knee of the curve that defines "loud enough" needs to be set, and it needs to be adaptive.

3. Formant-like spectral features are perceived independent of channel characteristics.

Techniques like CMN and Rasta are used in ASR to remove the effect of an overall channel transfer function (techniques like BSDCN try to do a better job of separating the channel transfer function effect on "loud enough" parts from the channel noise effects on "weak enough" parts, though in a domain where this separation does not work well). A key issue in all these techniques is the dynamics of adaptation, since channels change, sometimes quite rapidly; e.g., a user repeating a phrase more loudly can lead to a big effective change in spectral tilt (due not to the channel but to the source, but with similar perceptual effect).

4. Simultaneous masking: weak sounds with spectral energy near the spectral energy of strong sounds are less salient than weak sounds with spectral energy separated from that of strong sounds.

Spectral analysis schemes like MFCC, PLP, etc., have evolved with an auditory motivation to try to model critical-band simultanous masking; i.e., to not resolve in a spectral representation more detail than is resolved in the ear. Frequency warping in the LPC-Cepstral approach similarly tries to take advantage of what's known of auditory masking patterns. Generally these approaches have led to improvements.

5. Forward masking: weak sounds following strong sounds are less salient than weak sounds preceeding strong sounds; and weak sounds closely following strong sounds are less salient than weak sounds further separated from strong sounds.

ASR front ends almost always have a frame-by-frame approach with no account of forward masking, even though in psychoacoustics it is a huge and well-known effect. There is also backward masking, but it's relatively trivial by comparison. This time asymmetry is important but widely neglected.

6. Upward spread of masking: weak spectral features above strong ones (in frequency) are less salient than weak spectral features below strong ones (both in simultaneous masking and in forward masking). There is more upward spread at higher levels. Downward spread occurs, but is a much smaller effect.

ASR filterbank designs, etc., seldom account for the marked asymmetry of auditory filter shapes that lead to this frequency asymmetry in masking. Cochlea models usually do incorporate asymmetric passbands, though often in a very ad-hoc way.

An Aside on Cochlea Models:

The "Lyon's Cochlea Model" first hinted at in a term paper in 1978, first published in 1982 [1], with details trickling out at ICASSP 82-87 [1–6], and well described by Malcolm Slaney [7], was designed to account for all of these phenomena with a simple set of mechanisms (it

was also designed to provide a representation useful in accounting for pitch, which we'll cover later). The key combination of elements in the model and their importance are:

A: a cascade-structured filterbank: leads naturally to realistic asymmetric passbands with a low-complexity computational structure. Accounts for items 4 and 6: simultaneous masking and upward spread.

B: a half-wave detection nonlinearity: leads to an internal representation that correctly preserves all forms of temporal structure that lead to pitch and binaural effects (unlike what a square-law, full-wave, or envelope detector would do).

C: a coupled multi-channel AGC: provides for normalization of loud sounds and stable representation of silence, with dynamics designed to account reasonably for forward masking and channel insensitivity (primarily items 1, 2, 3, 5, plus a little downward spread of masking).

Features of later versions of the model, like a "cross-channel difference" (see Slaney's report [7]) lead to further improvements, especially in making a robust representation of formant peaks with good masking of noise via enhanced upward spread of masking. This mechanism and its noise suppression properties were independently discovered and well desribed by Shamma and his students in recent papers [8, 9].

One point on which the Lyon model has been perhaps deficient and Shamma has done better is item 1, the stabilization of the spectral shape representation of loud enough sounds. An alternate version of the Lyon model uses a sigmoidal rectification nonlinearity to better model signal detection by inner hair cells, as Shamma does, instead of the simple half-wave rectification nonlinearity, and leads to better results. I believe this will help the model work better with conventional ASR systems.

Item C, the AGC, is to me the single biggest contribution of my cochlea modeling work. It is also the most wide-open in terms of possible structural changes and parameter variations, and one of the least appreciated parts of my work among other auditory modelers.

The AGC can be viewed as a generalized and improved substitute for Rasta, replacing not only the filtering but also the logarithm, and thereby providing not just a solution for channel transfer function adaptation but also a robust way to account for loud vs weak sounds, forward masking, and adaptation on a wide range of relevant time scales.

7. Pitch as a carrier: sounds with a periodic pattern are more salient, especially in interference, than less structured sounds of a similar spectral envelope.

8. Binaural release from masking: sounds are much more salient if they arrive from directions different from the directions of interfering signals.

The ASR community has done lots of work on pitch and multi-microphone based enhancement, usually using adaptive linear filtering approaches (e.g. beam steering or adaptive antenna type techniques for multi-microphone, dereverberation, etc., and adaptive comb filters and such for enhancement based on pitch). These approaches have always had disappointing results.

As a step toward the longer term, it seems reasonable to extract pitch and binaural cues and

to track the speaker in pitch and space. How to do a good separation or enhancement given the cues and the tracking information is still a big open issue, but several ideas are around and have been studied some. An auditory approach here is very different from an adaptive linear approach. See the icassp paper and book chapter I did with Slaney for some ideas on using temporal information [10, 11].

Preliminary work on sound separation using binaural cues is in my ICASSP 83 paper [2] (and reprinted in the book "Natual Computation"); and on pitch based separation are Mitch Weintraub's ICASSP 84 paper [12] and Ph.D. thesis [13], and Duda et al [14].

9. Streaming: higher-level listening usually attends to a "stream" of sound input that is determined by a variety of self-consistent structural cues (spectral, temporal, pitch, binaural, etc.), ignoring other sounds that are not part of the stream of interest.

10. Gap-filling: when the interpretation of a stream implies the existence of a sound feature, but the feature is masked by other sounds (or is actually missing but would have been masked), then the feature is perceived as if present; i.e., it is "restored".

11. Gap sensitivity: when the interpretation of a stream needs a sound feature, but the feature is missing and not masked (e.g. a silent gap is inserted), then the interpretation is not accepted; i.e. silence is not as good a substitute for missing information as a masker is.

How to incorporate these effects into an ASR system is a fundamentally important question. In recent years a community of people working on "Computational Auditory Scene Analysis" has actually made some progress here. See especially the book from Martin Cooke's Ph.D. thesis [15] and the chapter derived from Mellinger's Ph.D. work [16].

References:

[1] Richard F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, May 1982.

[2] Richard F. Lyon, "A Computational Model of Binaural Localization and Separation", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston, April 1983.

[3] Richard F. Lyon, "Computational Models of Neural Auditory Processing", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, March, 1984.

[4] Richard F. Lyon and Niels Lauritzen, "Processing Speech with the Multi-Serial Signal Processor", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, March, 1985.

[5] Richard F. Lyon and Lounette Dyer, "Experiments with a Computational Model of the Cochlea", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, 1986.

[6] Eric P. Loeb and Richard F. Lyon, "Experiments in Isolated Digit Recognition with a Cochlear Model", Proceedings IEEE International Conference on Acoustics, Speech, and

Signal Processing, Dallas, 1987.

[7] Malcolm Slaney, "Lyon's Cochlear Model," Apple Technical Report #13, Apple Computer, Inc., Corporate Library, One Infinite Loop, Cupertino, CA 95104, 1988.

[8] K. Wang and S. Shamma, "Self-Normalization and Noise Robustness in Early Auditory Representations," IEEE Trans. Audio Speech 2(3):421-435, 1994.

[9] X. Yang, K. Wang, and S. Shamma, "Auditory Representations of Acoustic Signals," IEEE Trans. Information Theory 38:824-839, 1992.

[10] Malcolm Slaney and Richard Lyon, "A Perceptual Pitch Detector," Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, pp. 357-360, April 1990.

[11] Malcolm Slaney and Richard F. Lyon, "On the Importance of Time--A Temporal Representation of Sound," chapter 5 in Visual Representations of Speech Signals, M. Cooke and Steve Beet (eds.), John Wiley & Sons Ltd., 1992.

[12] Mitchel Weintraub, "The GRASP Sound Separation System", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, March, 1984.

[13] Mitchel Weintraub. "A theory and computational model of auditory monaural sound separation." Ph.D. Dissertation, Electrical Engineering Department, Stanford University, Stanford, CA, 1985.

[14] Richard O. Duda, Richard F. Lyon, and Malcolm Slaney. "Correlograms and the separation of sounds." In Conference Record. Twenty-fourth Asilomar Conference on Signals, Systems, and Computers in Pacific Grove, CA, Maple Press, 457-461, 1990.

[15] Martin P. Cooke "Modeling Auditory Processing and Organization," Cambridge University Press, 1993. (also Ph.D. thesis, University of Sheffield, 1991).

[16] David K. Mellinger and Bernard M. Mont-Reynaud, "Scene Analysis," chapter 7 in Auditory Computation, H. L. Hawkins et al. (eds.), Springer 1996.