# Speech Recognition Experiments with a Cochlear Model

**Richard F. Lyon**
Schlumberger Palo Alto Research
3340 Hillview Ave.
Palo Alto, CA 94304

**Abstract**

There are several ways that a computational model of auditory processing in the cochlea can be applied as the front end of a speech recognition system. For an initial round of experimentation, the fine time structure in the model's output has been used to do spectral sharpening, yielding a "cochleagram" representation analogous to a short-time spectral representation. In later experiments, fine time structure will be exploited for a more detailed characterization of sounds, and for sound separation.

So far, experiments have been done with only two words ("one" and "nine") spoken by 112 talkers, to limit the range of phonetic variation to simple voiced sounds, while providing a good sample of inter-speaker variation. The structure of the vector space of "auditory spectra" has been examined through vector quantization experiments, which yield a measure of information content and local dimensionality.

The inclusion of more dimensions of perceptual variation, such as pitch and loudness, in a speech front end representation is both an opportunity and a problem. Much larger vector quantization codebooks and more training data may be needed to take advantage of the extra information dimensions. A product-code approach and an improved algorithm for finding the nearest neighbor codeword are suggested to help cope with the problem and take advantage of the opportunity.

Preliminary recognition experiments using a single codebook per word and no time sequence information have shown a performance of about 97% correct one/nine discrimination for talkers outside the training set, and 100% correct for second repetitions from talkers in the training set. Further experiments are currently underway.

## 1 Introduction

Our experimental cochlear model has been most recently described in terms of its performance on simple "physiology" experiments [1]. Those experiments concentrated on the role of the AGC stages, which serve to partially normalize the output representation in the face of a wide dynamic range of overall amplitude and overall spectrum variations. The dynamics of the gain control process help to preserve perceptually relevant information about loudness and spectrum, emphasizing short-term changes.

The output of the model is regarded as a sequence of vectors in n-space, representing n-channel perceptual spectra. Silence maps to the zero vector, and perceptually louder sounds map to points further from zero. But detailed characterizations of this pattern space are difficult, due partly to its high dimensionality.

The number of important dimensions of variation due to phonetic and talker identity is an important issue in designing recognizers to work in this space, and is discussed in the next section. The following section discusses a set of recognition experiments, including comparisons with LPC. Finally, improved vector quantization techniques to work in this pattern space are suggested in the last section.

## 2 The Space of Cochlear Spectra

In the current version of the model, 92 bandpass channels are used to span a range of about 23 barks (about 100 Hz to 10 kHz). By modeling hearing, it is hoped that sounds will map into 92-space in such a way that a simple Euclidean distance in that space will correlate well with perceptual distinctions. Therefore, it is expected that a low-distortion vector quantizer designed to minimize mean squared Euclidean error will preserve most of the relevant information in a cochlear spectra. To explore this notion, codebooks of different sizes and distortions were constructed from various training corpora.

To make codebooks, a modified k-means algorithm was used. In each pass over the training data, new codewords were added to the codebook whenever the distortion to a training vector exceeded a desired distortion bound; at the end of a pass, each codeword was moved to the average of the vectors that were closest to it. Compared to a straight k-means with codebook size doubling, we found convergence to about the same rms distortion for a given codebook size, but in fewer iterations. Having maximum distortion as an independent variable is also useful.

The resulting data on codebook size *vs.* rms distortion and max distortion for a training corpus of 112 talkers saying "one" and "nine" are shown in Figure 1. The desired value of max distortion, such that reconstructed cochleagrams have clear and continuous formant and pitch tracks, is probably less than the lowest tried so far.

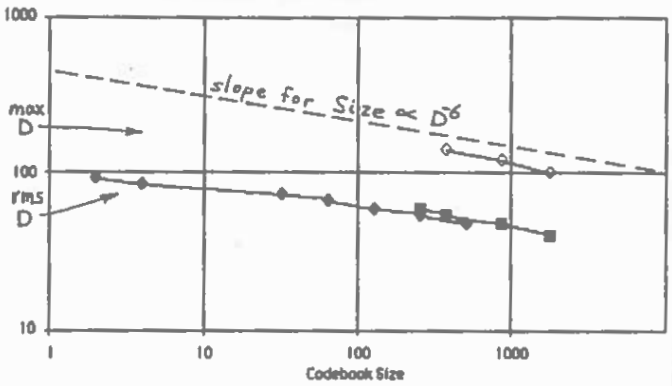


Figure 1: Codebook rms distortion (filled symbols) and maximum distortion (empty symbols) *vs.* codebook size.

The slope of the size *vs.* distortion curves (on a log-log plot) should reveal the dimensionality of the subspace that the codewords are packing into. Cutting the distortion by a factor of two will require a factor of sixteen in codebook size increase if there are four dimensions of variation to be covered.

The data show slopes corresponding to about 6 dimensions. Since the phonetic variation in the test corpus is quite small, much of this variation is probably due to talker differences. Since lower pitch harmonics are resolved in the spectrum, and loudness is not completely normalized out, these perceptually important dimensions contribute important dimensions of variation in the data that would not normally be seen in LPC and other common representations.

For the one/nine data, a codebook size of 1801 is barely adequate for high-fidelity coding of cochleagrams of the talkers in the training set. For the complete digit vocabulary, a codebook about five times larger would probably perform similarly. The distortion caused by using a codebook size of 383 is apparent in figure 2.

Based on these observations, it appears that representing a complete range of phonetic variation (eight or more dimensions), with reasonable fidelity would require a codebook size around 50,000 to 1,000,000. These sizes are far beyond normal practice in the speech recognition field, and require new techniques if they are to be useful.

## 3 Recognition Experiments with Cochleagrams and VQ Codebooks

Since training our existing recognizer [2] to use the cochlear spectrum pattern space will take considerable time, a much simpler test was undertaken first. Using the technique of Shore and Burton [3], a codebook was designed for "one" and another codebook was

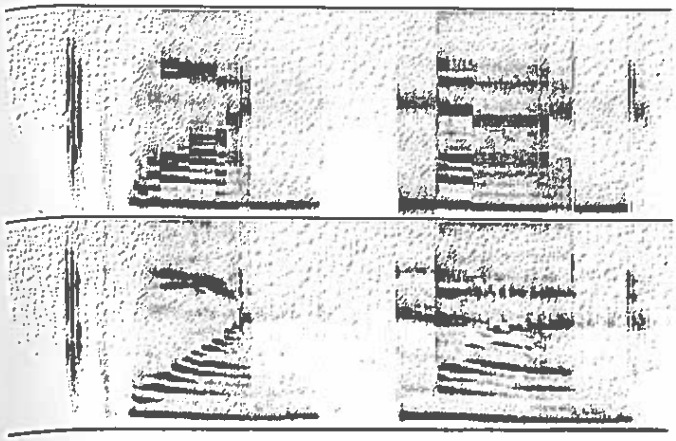

Figure 2: Cochleagram and vector quantized cochleagram of two digits by a talker outside the training set, with codebook size 383.

designed for "nine", using a single repetition of each word from each of the first 50 of the 112 talkers. Setting maximum distortion to 140 for both cases, the codebook for "one" reached a size of 261 and an rms distortion of 45.2, while the codebook for "nine" reached a size of 272 and a 5% higher rms distortion of 47.3.

Recognition proceeded by comparing quantization distortions (rms or total squared distortion) using the two codebooks, without compensation for the different codebook characteristics. No endpoint detection was done, so the generous amount of silence and noise at both ends of the words was included in the distortion measurements.

Testing on the second repetition of the same words from the training talkers led to no errors (in 100 trials). This result is encouraging, since this recognition technique has not previously been very successfully applied to speaker-independent or multi-speaker problems.

Testing on the other 62 talkers showed a serious bias: there were no misrecognitions of "one" as "nine", but ten misrecognitions of "nine" as "one" (5 on first repetition, 5 on second repetition, mostly from different talkers). Overall, on this speaker independent condition, there are 10 errors in 248 trials, or 96% correct. While this does not approach the performance of a good speaker independent isolated digit recognizer on the "one/nine" discrimination task, it is quite respectable for this simple algorithm.

Using order 11 LPC as a parameterization for comparison, with an Itakura distortion measure, we obtained at best 2 errors in 100 trials from talkers in the training set (98% correct), for various codebook sizes, and 14 errors in 248 trials on the other talkers (94.4% correct). Surprisingly, even very small codebooks (2 to 16 codewords) performed well with LPC, so it was decided to go back and try the cochleagrams with small codebooks.

With cochleagrams, it was found that for talkers in the training set, larger codebooks work best (sizes 32 and up gave no errors), but that smaller codebooks do a better job of generalizing to talkers outside the training set (size 32 was optimal with 7 errors in 248 (97.2% correct), while sizes 16 and 64 both were both slightly better than the initial large-codebook experiment, with 9 errors each. These differences may not be significant.

For every codebook size except size 2, the cochleagrams gave fewer errors than the LPC, usually by more than a factor of two.

## 4 VQ Algorithm Improvements

In spite of the encouraging results with small codebooks, it seems that to take full advantage of the information in cochleagrams with large talker populations will require very large codebooks. There are (at least) two alternative approaches to making very large vector codebooks practical. First, better fast quantization algorithms can be used to reduce the time cost. Second, codebooks can be constructed as product codes built from a small number of moderate-size codebooks.

Our present quantization algorithm takes advantage of the triangle inequality that applies to the Euclidean distance metric, so that codewords too far from a current best guess need not be examined; this unfortunately requires a table of $N^2$ inter-codeword distances, and so is impractical for much larger codebooks. The FN algorithm [4] uses a tree structure with a branch-and-bound search algorithm to take advantage of the same inequality with less stored information. Another approach which looks promising is to store the dual of the multi-dimensional Voronoi diagram [5] of the code vectors, so that each code vector is linked to its neighbors; in this case, when the current best guess is better than any of the neighbors, no further codewords need be examined. Using the last frame's quantization index as a first guess is very effective in these algorithms. In any case, the auxiliary data structures should be designed such that they are easy to modify when expanding or iterating the codebook.

The product code approach [6] is an alternative way to encode many bits of information per symbol with low distortion and small codebooks. The code space is the direct product of smaller codes, each of which encodes a separate part of the information in the original vector. In the simplest case, the original vector to be encoded is simply split up such that some components (*i.e.*, cochleagram channels) are used as a small vector in one codebook, and the other components are used with one or more other small codebooks. But other vector processing operations could also be used to try to separate the information more cleanly into feature vectors of lower dimensionality. For example, one process could attempt to capture pitch information, another could try to capture first formant information, etc. As long as these "feature extraction" processes don't lose information, the overall vector quantization distortion can be made as low as desired (even if quantizing sub-optimally by independently quantizing with each small codebook). If each feature detecting process captures only one or two important dimensions of variation, the resulting codebooks could be quite small. The structure imposed on the code space by the product code may also be useful in some kinds of recognition algorithms.

## 5 Conclusions

The cochlear model produces a spectral representation that captures important dimensions of speech signals. Preliminary experiments show that cochlear spectra lead to about 50% fewer errors in a very simple recognition technique, compared to LPC. Taking full advantage of the extra dimensions of information in cochlear spectra with a wide range of phonetic material and a wide range of talkers may yet require very large vector quantization codebooks or other techniques to extract the relevant features.

## 6 References

[1] Richard F. Lyon and Lounette Dyer, "Experiments with a Computational Model of the Cochlea," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Apr. 1986.

[2] Marcia A. Bush and Gary E. Kopec, "Evaluation of a Network-Based Isolated Digit Recognizer Using the TI Multi-Dialect Database," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Tampa, Mar. 1985.

[3] J. E. Shore and D. K. Burton, "Discrete Utterance Speech Recognition without Time Alignment," *IEEE Trans. Inform. Theory* **IT-29**, pp. 473–491, July, 1983.

[4] K. Fukunaga and M. M. Narendra, "A Branch and Bound Algorithm for Computing *k*-nearest Neighbors," IEEE Trans. Computers, **c-24**, pp. 750–753, 1975.

[5] D. T. Lee and Franco P. Preparata, "Computational Geometry—A Survey," IEEE Trans. Computers, **c-33**, pp. 1072–1101, 1984.

[6] John Makhoul, Salim Roucos, and Herbert Gish, "Vector Quantization in Speech Coding," Proc. IEEE, **73**, pp. 1551–1558, 1985.

Supported by DARPA contract N00039-85-C-0583.