REPORT:

N. S. Jayant *is head of the Signal Processing Research Department at AT&T Bell Laboratories in Murray Hill, New Jersey. He is responsible for research in speech and image processing with applications to coding, communications, and recognition. Mr. Jayant joined the company in 1968 and has a Ph.D. in electrical communications engineering from the Indian Institute of Science, Bangalore, India.*

EFFECTIVE NUMBER OF KEYS IN A VOICE PRIVACY SYSTEM BASED ON PERMUTATION SCRAMBLING

Introduction

In a voice privacy system that is based on permutation scrambling of N speech elements, the total number of permutation *keys* is $N!$. It is generally acknowledged that this number is an exceedingly optimistic indicator of cryptanalytical strength. Because of the redundancy in the speech signal and the intelligibility of imperfectly inverse-permuted speech, an eavesdropper often needs to try only $K << N!$ random permutations before *breaking* the speech code.

A general estimate of the *effective* number of keys K in speech permutation is difficult to define or evaluate. Therefore, this paper provides an order-of-magnitude estimate that applies to a pure-listening type of attack in a specific but important class of privacy systems. These systems are based on permutations of a time-frequency speech matrix whose duration is typically several tens of milliseconds, and whose elements are speech samples from each of about four or five contiguous frequency subbands.[1] Even within this specific context, because of the highly simplified nature of the speech model, this paper's results provide a semiquantitative perspective rather than definitive numbers for key strength.

The Speech Model. For the analytical estimate of K, we reflect the redundancy in (voiced) speech through an idealized integer P, the number of pitch periods (with identical waveforms) in the speech matrix as the input to the permuter. To reflect that imperfectly permuted speech is intelligible, we use another idealized parameter F, which is the fraction of speech elements that need to be exactly *in place* for speech information to be available to an eavesdropper in a pure-listening type of cryptanalytical attack.

The two-parameter model based on P and F is not a complete or canonical one. Instead, it is a first attempt to provide a structured attack to a long-standing problem, that of evaluating K. We believe that this model, imperfect as it is, leads to some interesting perspectives on analog voice privacy. Further, we hope that the introduction and use of this model may trigger work of greater generality and rigor.

The parameters P and F are described more completely later. The effective number of keys K—expressed as a function of N, P and F—shows the general looseness of the nominal estimate $N!$. Numerical calculations—for the examples of $N = 128$; $P = 1, 2, 4,$ and 8; and $F = 1.0, 0.5, 0.25,$ and 0.125—show a dramatic reduction of K for the extreme case of $P = 8$ and $F = 0.125$.

The value of F that is meaningful in speech decryption (by pure listening) should depend highly on context. Thus, it is difficult to justify the use of a specific value of F as a global model. Further, the parameter F does not reflect the dependence of intelligibility on the way in which correctly deciphered samples may be clustered or separated in time and frequency. However, on the average, we propose that $F = 0.5$ corresponds roughly to a signal-to-distortion ratio, $[10 \log_{10} (F/1 - F)]$, of 0 dB, which, in turn, is a reasonable threshold for speech intelligibility.

To make our interpretations conservative, we will look at F values that correspond to signal-to-distortion ratios that are <0 dB ($F < 0.5$). We will note that F values of 0.25 still lead to very large values of K if N is large enough.

On the other hand, the dramatic reduction of K when F is 0.125 shows the importance, for extremely high levels of security, of a dynamic permutation procedure where cryptanalytical strength derives from adequately frequent time variation of the permutation key, rather than the key's cardinality as expressed by $N!$.

The unsophisticated eavesdropper assumed in this paper can figure out a useful random permutation for a given input segment after K brute-force experiments. But with a time-varying scrambler, the eavesdropper has to repeat this quest to decipher a speech sequence that has been permuted with a different key.

The Parameter *P*

We use a highly simplified input model where the block of N speech elements to be permuted contains exactly P identical segments, each with N/P elements. Further, we assume that the speech segments are split into contiguous frequency bands (Figure 1), so that the input to be scrambled is a succession of two-dimensional blocks of time-frequency samples.[1] For simplicity, we assume that P and N/P are integers, with $P \geqslant 1$. This part of the model is inspired by a highly periodic voiced-speech example.

In the model, corresponding elements in successive pitch periods are identical in amplitude. The assumption of P identical sub-blocks gives us a maximum-redundancy model that, in turn, will lead to a maximally conservative estimate of K (for a given value of F) from the viewpoint of the intended user of the system.

In the special case of time-frequency segment permutations, where speech is split into contiguous frequency bands (such as the four frequency bands in Figure 1), the strongest kind of redundancy is indeed that reflected by the distant-sample correlations of a periodic (subband) waveform. Adjacent sample correlations in subband waveforms are much less substantial than the adjacent sample correlations in full-band speech. They are certainly much weaker than the perfect distant-sample correlations in our idealized model.

In summary, then, to incorporate speech redundancy in the analysis, the proposed model uses the single parameter P. The results of such an analysis are expected to reflect the key strength of a subband permutation privacy system.[1] For privacy systems that are based on the permutation of a full-band speech signal, the analysis will also have to include parameters that characterize adjacent sample correlations.

The Parameter *F*

We use another idealized parameter, F, to model the fact that imperfectly inverse-permuted speech is intelligible. The parameter F is defined as the fraction of speech elements that must be *in place* for speech information to be available in a pure-listening cryptanalytical attack. In the model of Figure 1, where each speech amplitude is repeated exactly P times,

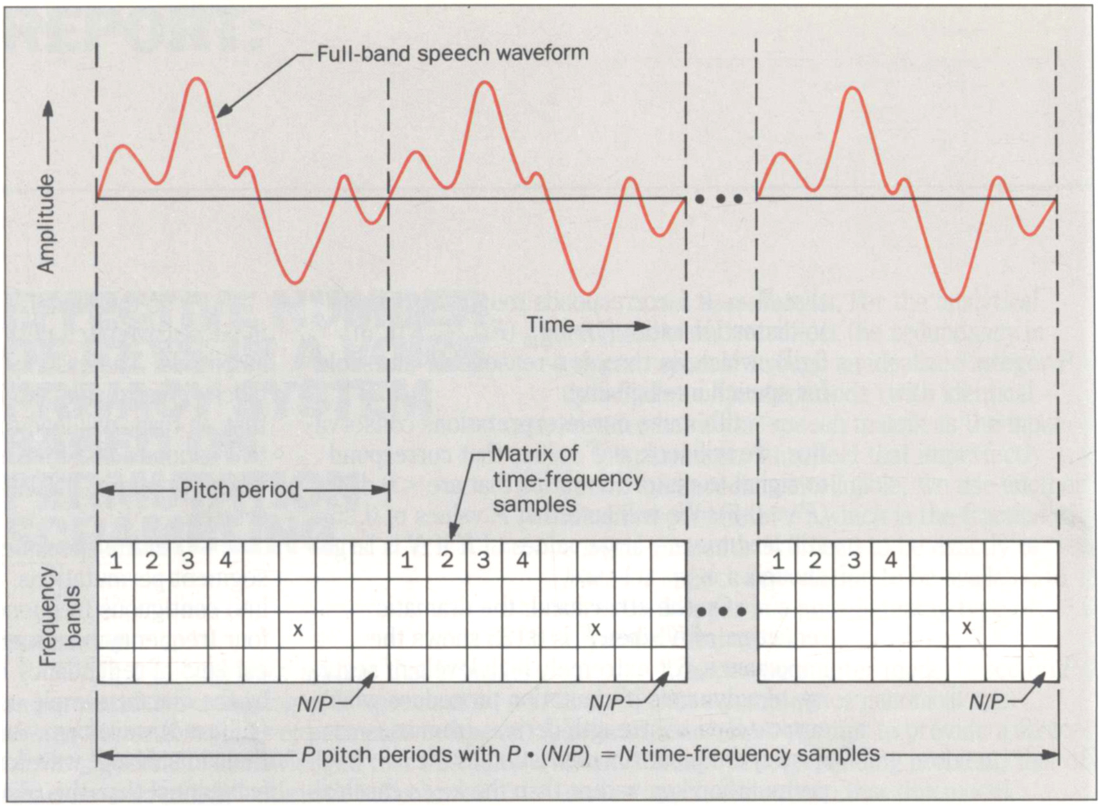
Full-band speech waveform
Amplitude
Time
Pitch period
Matrix of time-frequency samples
Frequency bands
1
2
3
4
x
N/P
P pitch periods with P • (N/P) = N time-frequency samples

Figure 1. Simplified model of speech input in time-frequency sample permutation. The input buffer contains exactly *P* pitch periods, each with *N/P* samples. With *B* frequency bands ($B = 4$ here), each subband has *N/PB* samples in a given pitch period. The *P* samples in a given time-frequency position *X* (where $X = 1, 2, \ldots, N/P$) are equal in amplitude.

a speech element will be in place if, after permutation, it occupies either its original place or any one of $(P - 1)$ corresponding places in the $(P - 1)$ repetitions of the basic speech matrix.

F is idealized in that it depends highly on context. In particular, it may also depend on the parameter P. It is also a simplified description of the perceptual process, because the intelligibility of partially inverse-permuted speech depends on several things:

- What fraction of elements are in place?
- Which particular elements are in place?
- What exact way are these elements positioned, separated, or clustered in time and frequency?

The effects of severe clusters and other pathological configurations cannot be reflected in our simplified theory. But we believe that these effects will become less significant as the block length N becomes larger. The parameter F will consequently be more meaningful for large N.

This paper's results refer to the case of large N, in particular to $N = 128$.

The Probability *p(N, F, P)*

In this paper, we focus on the probability p of encountering (in a search through the dictionary of $N!$ keys) a permutation key in which a compromising amount of speech information is available. This probability will be a function of N, P, and F. For large N, it will have three interpretations:

1. If the transmitter can eliminate the fraction p of compromising permutation keys (unrealistic in general), the total number of useful keys left in its dictionary is $K_1 = N!(1 - p)$.
2. If the transmitter uses random permutations that are selected from the entire unedited dictionary of $N!$ keys, the number of good keys it uses before encountering a bad one (a compromising key) is about $K_2 = p^{-1}$. The expected number of bad keys is $M = N!p$. If the locations of these keys are uniformly dis-

tributed in the range 1 to $N!$, then $N!/(M + 1)$ is the expected value for the first time a bad key is encountered (from the theory of extremal statistics). With $M = 1$ (the classical case of a single bad key), this number is the familiar *first-encounter* result $N!/2$. With $M = N!p$ and $N!p >> 1$, the expected value of $N!/(M + 1)$ is very close to $1/p$.

3. Assume that the transmitter is using a good permutation (this happens hopefully with a high probability $1 - p \sim 1$). Then, an eavesdropper who tries to understand a speech block, by trying random *inverse* permutations followed by listening, will succeed (i.e., encounter the first compromising inverse permutation) after a number of experiments whose average value is about $K_3 = p^{-1}$. (The product of two random permutations is also a random permutation, so by definition $K_3 = K_2$.)

For the example of an unsophisticated transmitter followed by an unsophisticated eavesdropper, the selections of permutations and inverse permutations are both random. In such a system, the effective number of keys is

$$K = K(N, F, P) = K_3 = [p(N, F, P)]^{-1} \quad (1)$$

Table I. Order-of-Magnitude Estimates for Number of Keys

	P			
F	1	2	4	8
1.0	3.8×10^{215}	2.0×10^{196}	2.6×10^{171}	7.9×10^{141}
0.5	3.4×10^{89}	6.4×10^{74}	1.1×10^{58}	4.5×10^{40}
0.25	6.9×10^{35}	3.0×10^{27}	8.5×10^{18}	6.8×10^{10}
0.125	5.3×10^{13}	3.1×10^{9}	3.1×10^{5}	1.6×10^{2}

NOTE: These estimates on the effective number of keys $K(128, P, F)$ are for the conservative model of Figure 1. (The speech is permuted within blocks of $N = 128$ time-frequency elements. Values of $P = 1$ and 8 typically imply pitch frequencies of 40 and 320 Hz.[1] A value for $K = 10^4$ implies 10 hours of code breaking, if each subexperiment that consists of permutation followed by listening needs 3.6 seconds.)

The Expression for K

The problem is purely a combinatorial one at this point. We have an input of N elements that belong to N/P classes, and each class has P identical elements. We want to derive an expression for the total number of distinct permutations in which exactly NF elements are in place. A speech element is *in place* if it occupies the position that belonged originally to any of the P elements in its class, including itself.

In each case, the effects of $P > 1$ and $F < 1$ are to make the number of desired permutations K less than the nominal total of $N!$. This, in turn, means that an eavesdropper, who relies on a sequence of subexperiments that each consist of random permutation followed by listening, has to go through only K rather than $N!$ subexperiments on the average before realizing intelligible speech. Clearly, if $F = 1$ and $P = 1$, then $K = N!$.

The probability of getting *at least NF* of N elements in the correct locations in a random permutation is

$$p = \sum_{n = NF}^{N} p_n \quad (2)$$

where p_n is the probability of *exactly n* correct placements. Expressions for p_n can be determined with methods described by Riordan,[2] who solved the equivalent problem of obtaining n matches in two card decks that each contain

P repetitions of N/P card types. Computer programs have been written to obtain values of p_n, and these values have been used in eq. (2) to obtain exact expressions for the cumulative probability p in eq. (2). Reciprocals of these values of p give corresponding values of K_3, as in eq. (1).

Table I shows rounded values of $K(N, P, F)$ for $N = 128$; for $P = 1, 2, 4$, and 8; and for $F = 1, 1/2, 1/4$, and 1/8. The table shows a dramatic reduction in the estimated key strength K, if $F << 1$ and $P >> 1$.

The value of the critical fraction F is hard to quantify, as discussed earlier. We feel that F cannot be too small, because the NF samples that are in place have to be perceived against the heavy masking that results from the $N(1 - F)$ noisy samples [leading to a signal-to-distortion ratio of $10 \log (F/1 - F)$ dB]. We believe that values of F of about 0.25 (signal-to-distortion ratio of -5 dB) may represent a conservative threshold for intelligibility. We also believe that $F = 0.125$ (signal-to-distortion ratio of -8.5 dB) may represent a realistic threshold only in specific cases.

A value of $P >> 1$ is realistic in typical implementations of a time-frequency scrambler;[1] the extreme value of $P = 8$ in Table I represents a pitch frequency of 320 Hz. (In Figure 1, $N/4P = 128/32 = 4$ samples at a subband sampling rate of 1280 Hz.)

The values of K for $F << 1$ and $P >> 1$—in particular, for $F = 0.125$ and $P = 4$ or 8—may indicate the desirability, for very high security, of a dynamic system that derives cryptanalytical strength from adequately frequent key variation rather than the nominal key cardinality as expressed by $N!$. This observation may be particularly valid for permutation scrambling of a full-band speech signal, a situation characterized by adjacent sample correlations that could reduce the key strength to lower values than those predicted for the subband signals of Figure 1. Consider the subband scrambler that has been the focus of this paper and another article.[1] If we assume $F = 0.25$ is a realistic intelligibility threshold and 3.6 seconds is an illustrative value of the time per permute-listen subexperiment, then Table I shows that even the extreme case of $P = 8$ implies a *code-breaking* time of 68 million hours!

Acknowledgment

I would like to thank J. A. Reeds and P. G. Doyle for pointing out that the probability p in eq. (2) had an exact solution. I am grateful to M. Merritt for his incisive comments on an earlier version of this paper.

References

1. R. V. Cox et al., "The Analog Voice Privacy System," *AT&T Technical Journal*, Vol. 66, No. 1, January/February 1987, pp. 119-131.
2. J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley & Sons, New York, 1958.

(Manuscript received May 2, 1986)