REPORT:

Authors:
***Martha Birnbaum** and **Larry A. Cohen** are members of the technical staff in the Business Systems Applied Research Department of AT&T Information Systems. **Frank X. Welsh** is a senior technical associate in that department. Ms. Birnbaum is working on automatic speech recognition. She has a B.A. in German from Bryn Mawr College and an M.A. and Ph.D. in linguistics from Brown University. Mr. Cohen is involved in applied research in artificial intelligence technologies. He received a B.S. in physics from Yale University and an M.A. and Ph.D. in physics from Harvard University. Mr. Welsh is responsible for new techniques in speech and signal processing hardware and software. He received a B.S. in electrical engineering from the Pennsylvania State University.*

A VOICE PASSWORD SYSTEM FOR ACCESS SECURITY

Introduction

A voice password system for access security using speaker verification technology has been designed for use over dial-up telephone lines. The voice password system (VPS) can provide secure access to telephone networks, computers, rooms, and buildings. It also has application in office automation systems, electronic funds transfer, and "smart cards" (interactive computers embedded in credit-card-sized packages). As increasing attention is focused on access security in the public, private, and government sectors, the voice password system (Figure 1) can provide a timely solution to the security dilemma.

The VPS uses modes of communication available to almost everyone (the human voice and the telephone). A user calls the VPS, enters his or her identification number (ID) by touch-tone telephone, and then speaks a password. This is usually a phrase or sentence of about seven syllables.

On initial calls, the VPS creates a model of the user's voice, called a reference template, and labels it with the caller's unique user ID. To gain access later, the user calls the system, enters the proper user ID, and speaks the password phrase. The VPS compares the user's stored reference template with the spoken password and produces a distance score.

If the score is below a preset threshold, the VPS decides that the caller and the person who created the reference template are the same person and grants the caller access. If the score is above the threshold, the caller is considered an imposter and access is denied.

System Design

AT&T Information Systems has built a prototype voice password system. The main components of the system are (see Figure 2):

- One or more telephones
- One or more custom-designed voice verification unit circuit boards that can operate simultaneously
- A host computer for reference template storage and system administration
- A system administrator's console.

In Figure 2, an AT&T 3B2 microcomputer is the host computer. However, any UNIX®-based computer or PC (for example, the AT&T UNIX PC) can be used for the storage/administration function.

Verification telephones need not be located near the voice password system. They can be at a user's desk or home for remote access to a computer. Or, they can be near an entry portal for site access. A relay, controlled by the voice verification unit, activates a circuit that opens an entry portal to authorized users. Access to computers or telephone networks is controlled by the host with instructions from the voice verification unit.

The workhorse of the VPS is the voice verification unit, a single circuit board. This multifunction unit (Figure 3) does the following:

- Answers a call and processes the caller's touch-tone user ID number
- Provides speech prompts to the caller
- Requests the user's reference template from the host database
- Gathers the voice password and extracts digital features from the speech
- Compares these features with the reference

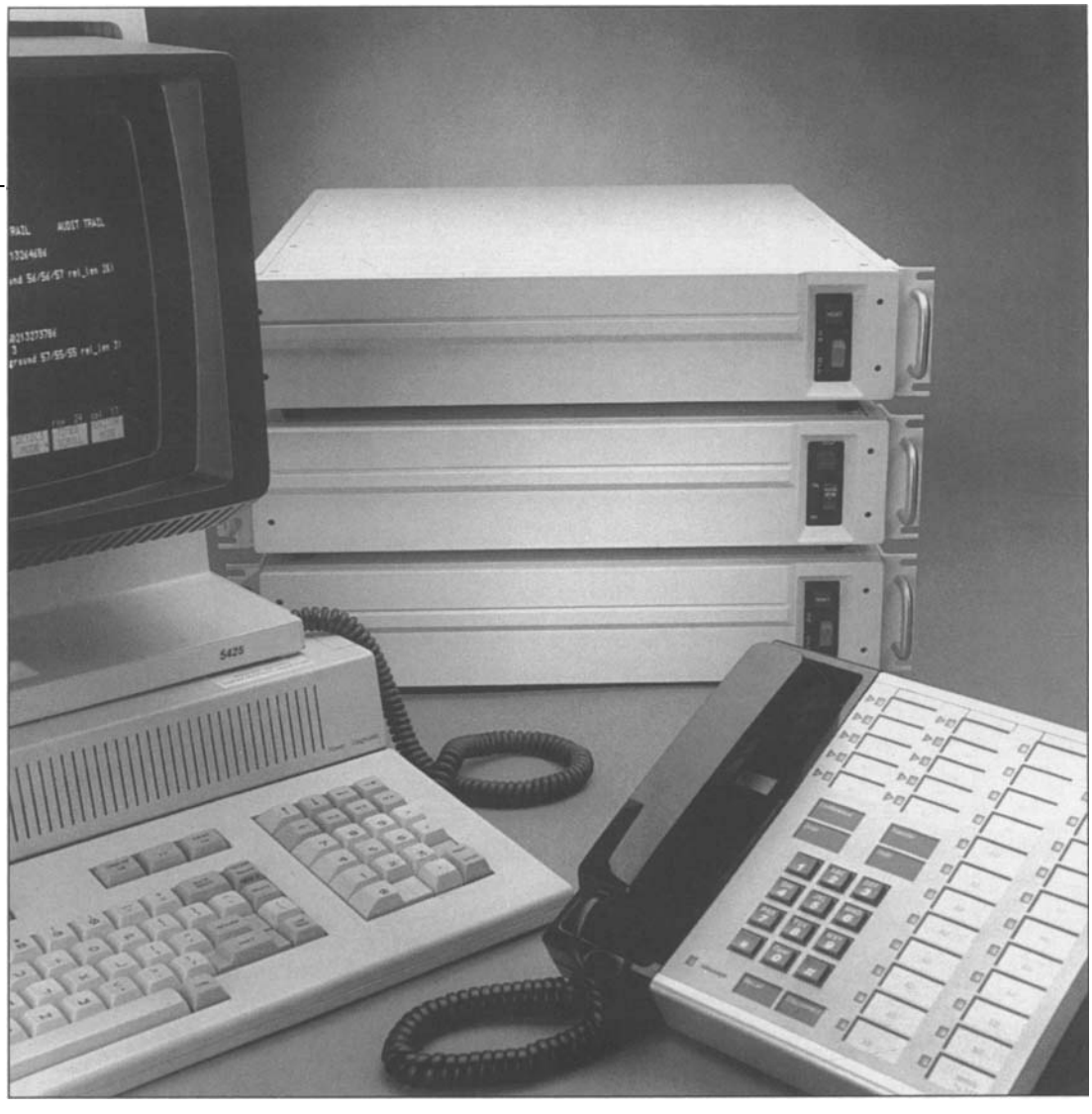

Figure 1. The voice password system consists of the voice verification unit (three units are multiplexed here), a computer, an administration terminal, and a telephone.

template
- Decides whether the comparison is close enough and notifies the host
- Updates the user's reference template and sends it back to the host database.

The telephone interface answers a call and sends incoming speech to a coder/decoder (codec), where it is digitized and sent to an AT&T 310D digital signal processor (DSP). The DSP performs preliminary speech processing functions on the digitized signal, and passes a greatly compressed amount of information to the central processing unit (CPU).

The CPU performs or supervises the remaining operations in the verification task. It matches the incoming utterance against the user's claimed reference, makes the accept/reject decision, and handles overall administration and timing for the board.

The AT&T 439B speech synthesizer chip creates high-quality prompts for the caller. Such prompts are phrases stored in nonvolatile memory chips or EPROMs (erasable, programmable read-only memory).

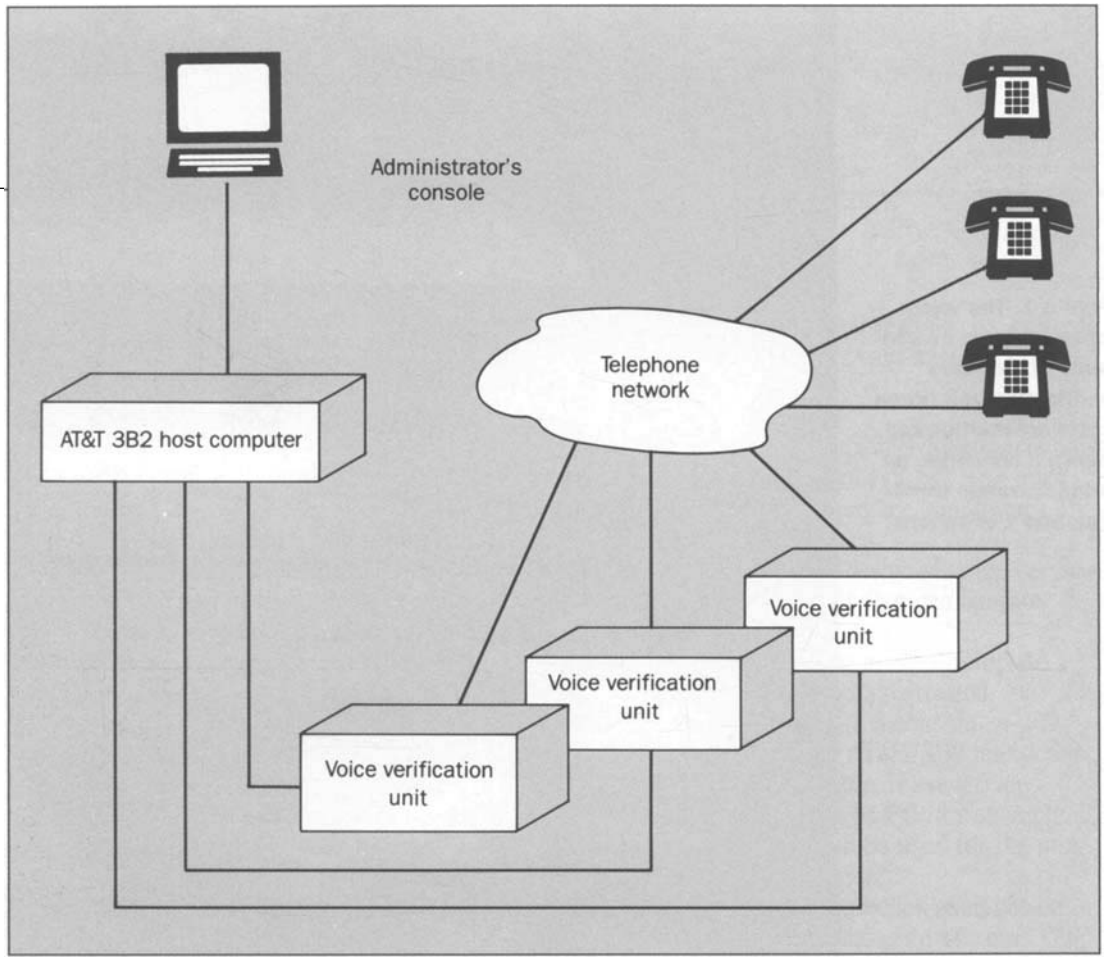

Figure 2. A block diagram of the voice password system.

Speech Processing

Speech entering the 310D digital signal processor is continuously processed on a frame-by-frame basis. Frames are spaced at 15-millisecond (ms) intervals, and overlap with a 45-ms duration.

For each frame, the natural redundancy of the speech signal is drastically reduced by extracting a set of features that characterize the important aspects of the signal, namely the short-term energy and the spectrum. The spectrum reflects not only the phonetic content of the speech but also the shape of the vocal tract that produced it. Thus, spectral representations may reflect differences among speakers.

The first step in feature extraction is to compute the autocorrelation of the incoming signal. The autocorrelation coefficients are then modified by simulating the addition of white noise to reduce the differences between noisy long distance lines and clear local lines. This technique, which limits the dynamic range of the speech to 30 dB, has been shown to improve the performance of the verification system.

The modified autocorrelation coefficients are transformed into linear predictive coding (LPC) coefficients. LPC coefficients represent, in a small number of bits, a spectrum of the voice.

As LPC feature extraction occurs, the central processing unit transforms the LPC coefficients to cepstral coefficients. The cep-

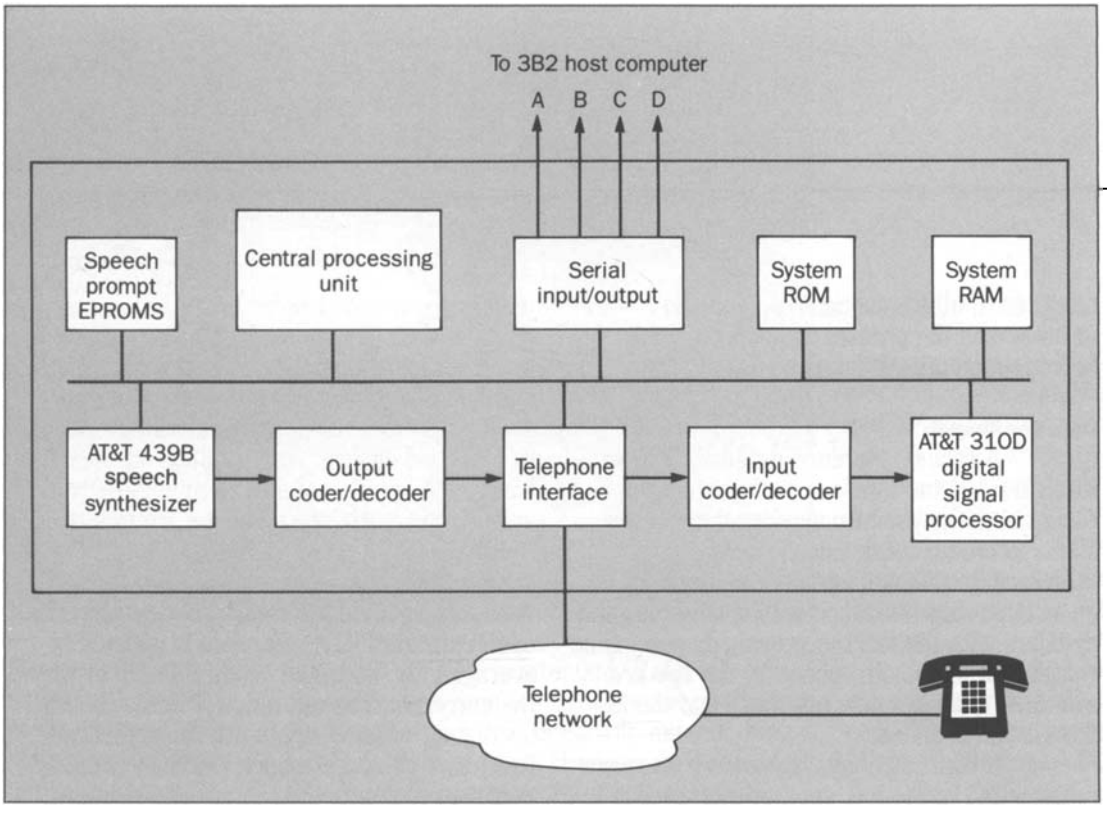

Figure 3. The voice password system custom board provides voice prompts, input speech analysis, voice pattern matching, and control.

strum is the inverse Fourier transform of the log power spectrum. For computational reasons, we use the LPC-derived cepstrum, which is a transformation from the LPC coefficients. Although LPC has been used for many years in automatic speech recognition and synthesis, cepstral coefficients have been found to differentiate between speakers more accurately than LPC coefficients while still characterizing vowels properly.

The cepstral coefficients are further modified by subtracting the mean cepstral values over the utterance from each 15-ms frame of speech. This technique, called channel normalization, has the effect of dramatically reducing filtering effects of the long-distance channel.

After feature extraction, the CPU makes an endpoint detection. That is, it locates the beginning and end frames of the password phrase. Endpoint detection is based on signal energy in relation to several energy thresholds computed from a constantly updated estimate of background noise level.

Pattern Matching

Once a feature set for a given utterance has been computed, it is matched with a previously generated reference pattern using *dynamic time warping* (DTW). This match, a variant of dynamic programming, accounts for timing differences among repeated utterances of the same phrase. DTW works by nonlinearly shrinking or expanding the speech to "line up" the syllables of the utterance and the reference.

The DTW match produces an absolute distance score that is an estimate of how different the utterance is from the stored reference. The score, which would be lower for correct utterances, is used to evaluate the identity claim. If the syllables of an utterance line up well with those of the reference, the score then represents entirely phonetic, or spectral, dif-

ferences. If utterance and reference do not line up well, or if the phrases are different, the score is dramatically increased.

User Enrollment

A robust reference template that accurately models the user's speech patterns is essential to the performance accuracy of the VPS. To create a reference template, one utterance is collected and used as the reference. A second utterance is then collected and matched with the reference using dynamic time warping. If the match succeeds, the two are averaged to form a new reference and the initial session is ended.

If the match fails, the newer utterance replaces the reference, another utterance is collected, and the process is repeated. If four utterances are collected without two in a row matching, the session is ended without any reference being stored. The speaker is instructed to hang up and call again to repeat the training session.

Immediately following reference creation, the user makes a series of calls to the system. These *enrollment calls* accustom the user with the system and make the reference template representative of the user's speech patterns.

Evaluating the Identity Claim

A user's password phrase is compared to the reference of the true speaker using dynamic time warping. The resultant score is compared to a preset threshold. If the score for a user's password attempt is below the threshold value for the true speaker, the identity claim is accepted.

To reduce the effects of perturbations of the utterance (coughs, door slams, etc.), the user is given two chances to speak the password phrase. This procedure helps true speakers more than it helps imposters. A true speaker rejection is often the result of poor speaker performance, and is usually remedied by a second chance. An imposter, on the other hand, is generally not able to mimic the true speaker any better on a second attempt than on the first.

Both the reference and the threshold value are updated after each successful verification attempt. The reference is updated by averaging the accepted cepstral features with the corresponding reference features. Moving averaging, which weights utterance features less heavily than reference features, is used so that the reference will contain contributions from many utterances. Because a person's speech varies over time, updating ensures that the reference will continually adapt to the user's current speech patterns.

Figure 4 shows the history of a typical user's distance scores. After the enrollment period, scores become gradually lower, indicating that the reference is "learning" the range of variation in the user's voice. The threshold is updated by averaging the current score plus a constant "bias" factor with the previous threshold. Updating is illustrated in Figure 4 by the downward slope of the threshold.

The bias is a system-wide parameter controlling the balance between true speaker rejections and imposter acceptances. Reducing the bias makes it more difficult for both true speakers and imposters to be accepted. Increasing the bias raises the threshold, making it easier not only for true speakers, but also imposters, to be accepted.

In practice, the VPS administrator sets the bias value according to the application

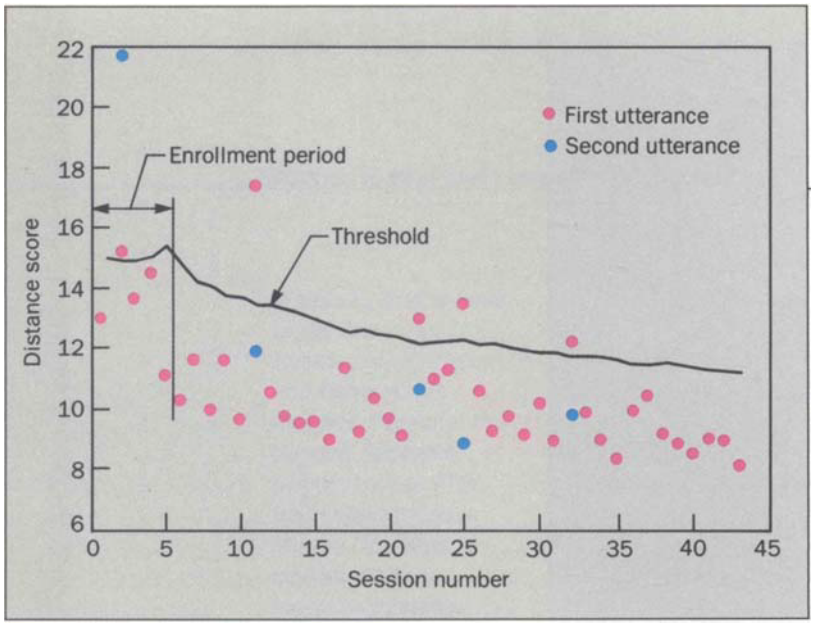

Figure 4. Profile of a typical caller's distance scores. The distance score is an estimate of how different the caller's spoken password is from the stored reference. When the user is rejected (i.e., distance score is greater than the threshold), a second chance is given.

for which the VPS is intended. If, for example, the application requires tight security, reducing the bias to satisfy this need will, indeed, deny access to all imposters. However, in doing so, it will also deny access more often to some true speakers.

System Performance

An evaluation of the performance of the voice password system was conducted in December 1985. Twenty-one speakers (twelve males and nine females) made a minimum of 40 calls to the VPS during working hours over a ten-day period. All speakers used the same password phrase. (In practice, password phrases can be secret and different for each user, making it far more difficult for an imposter to be accepted as a true speaker.) This procedure provided about 926 calls. Sixteen speakers called locally (710 calls) and five called long distance (216 calls).

To collect imposter utterances, 23 males and 22 females who were not part of the original population were asked to make two calls (four utterances) each to the VPS, using the same password phrase as the true speaker population. These were then matched off-line against all the true speaker references that were of the same sex.

The results of the performance evaluation are shown in Figure 5. The crossover plot consists of simultaneous plots of Type I and Type II error rates as a function of the bias. A *Type I* error, the rejection of a true hypothesis, represents the rejection of a true speaker. A *Type II* error, the acceptance of a false hypothesis, here represents acceptance of an imposter. The operating bias of the performance evaluation was 1.4 (as labeled in Figure 5).

At the operating bias, the true-speaker rejection rate (Type I) was 3.1 percent. There were no imposter acceptances (Type II). Because of the relative flatness of the Type I "tail," our operating bias of 1.4 is close to a point of diminishing returns. Increasing the bias would not significantly reduce true speaker rejections but would soon start to increase imposter acceptances.

The *equal error rate,* the intersection of Type I and Type II curves, is frequently used to characterize verification systems. For this performance evaluation, the equal error rate was 1.9 percent. An earlier test of approximately the same size gave similar results.

Eighty-five percent of the errors causing the high tail of the Type I curve were caused by endpoint-detection errors attributable to circuit noise on the board used in the performance evaluation.

This problem has since been remedied, allowing us to project the operating characteristics of the voice password system to be 1.4-percent Type I error, and 0.4-percent equal error rate. These values are adequate for most security applications.

Normal variations in speakers' voices are expected. This leads, of necessity, to some

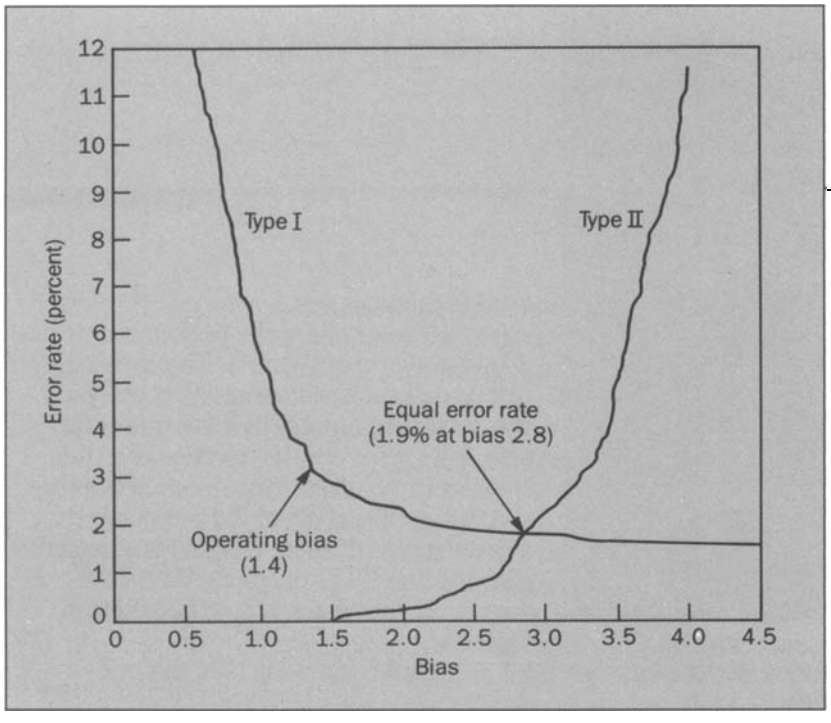

Figure 5. A performance evaluation conducted in December 1985 measuring the speaker accept/reject error rate of the voice password system.

level of true speaker rejection. This is the cause of the remaining errors. Because of this intrinsic variation in human speech, no voice verification system can give 0-percent true speaker rejection at any bias reasonable for high-security applications.

Conclusion

Several voice verification units have been built and distributed to various groups within AT&T for testing and evaluation. Possible applications of voice password technology are being studied for access to telephone features via PBX's, for access to telephone networks and computers, and in digital switches.

Additional applications being investigated are computer room access, office automation, and telephone credit card verification.

(Manuscript received July 17, 1986)