Speaker Recognition from an Unknown Utterance and Speaker-Speech Interaction

R. L. KASHYAP, MEMBER, IEEE

Abstract—We are interested in determining whether the given utterance comes from a member of a given speaker group or an imposter. If it is the former, we are interested in determining the identity of the speaker. The only knowledge available is a set of known utterances from the given group of speakers. The given utterance is manually divided into phonemes without necessarily ascertaining the identity of phonemes. Using statistical decision theory, we will develop various types of tests for speaker verification and identification using only one phoneme segment or the entire utterance. We will consider related problems such as the methods of clustering speakers to aid speaker verification, the optimal choice of phonemes for speaker recognition. Next we consider the role of speaker variability in speech recognition and recognize its complementarity to the problem of optimal choice of phonemes for speaker recognition. We illustrate the efficacy of the various methods developed here by considering the speaker and speech identification problems with three speech data bases.

I. Introduction

WE ARE interested in the problem of recognizing a speaker from an unknown utterance emitted under uncontrolled conditions. This problem is studied in the broader perspective of the relationships between the speaker and the phoneme in producing the final utterance. The utterance could be a word or a phrase. As is well known, there are two aspects of speaker recognition. First, there is the problem of checking whether the utterance is from a person in a known group or from an imposter—the so-called speaker verification problem. Second, we may want to determine the identity of the speaker from the utterance given that the utterance comes from one of the persons from a known group—the so-called speaker identification problem. We will consider both problems. The only knowledge we have about the speakers in the group is a set of waveforms of known words and phrases uttered by each speaker. By definition, we cannot have any information about the imposter. Both of the problems mentioned above are of considerable importance in many cases of practical interest, such as white collar criminal identification, development of voice-actuated secure identification systems for the users of a facility such as the computer, etc.

The literature on talker verification with known utterances and identification is considerable. There is a good review of the literature in [1]. Some of the recent contributions are in [2]-[4]. The general conclusion is that speaker identification can be successfully done in a variety of ways using widely differing sets of features. It is difficult to compare different methods because of the lack of any systematic theory of the interaction of the speaker and speech in producing the final utterance. The success of the various methods only shows the large amount of redundancy in the utterance for the purposes of speaker identification. There is no way of comparing the various methods except on the basis of their recognition performance. The number of such studies to date is very small.

Since we are interested in the behavior of speaker and speech interaction, we will develop our recognition schemes and the associated theory in terms of phonemes, i.e., the given utterance is manually divided into phonemes without necessarily ascertaining the identity of phonemes. The phoneme in each segment is only known to belong to a certain set Ω. We will initially consider methods for speaker verification and identification based on a single segment of the utterance and extend them so as to utilize the information available in all the segments of the utterance. Next we want to ascertain the role of faulty segmentation on the performance of the speaker recognizer. We may note that loss of accuracy in the speaker identification due to the ignorance of utterance is not substantial.

All of the above methods are based on one particular (statistical) hypothesis of speaker-phoneme interaction. We consider alternative hypotheses as well to show the effectiveness of the particular hypothesis.

These methods are impractical while dealing with large groups of speakers. In such cases, we have to divide the speakers into subgroups. We will consider methods of clustering the speakers.

We will also investigate whether certain classes of phonemes like nasals or vowels are significant from the point of view of speaker verification, i.e., given the choice of the test word for utterance, it should have phonemes so as to reduce the probability of error. Finally, we will show that even when we are interested in the determination of speech (i.e., phonemes) of an utterance from an unknown speaker, recognition methods which emphasize the speaker variability give better preference than those which do not.

The methods of this paper are expressed in terms of phonemes because phonemes can be considered to be the basic units of speech. However, in recent literature some investigators [6] have vigorously expressed that the "distinctive features" should be regarded as the basic units and phonemes should be regarded as combinations of these units. Hence, in the beginning we tried to develop speaker recognition methods

Manuscript received February 9, 1976; revised July 13, 1976. This work was supported in part by the National Science Foundation under Grant Eng 74-17586 and in part by the Air Force Office of Scientific Research under Grant 74-2661.

The author is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

which use the distinctive features. The results are considerably poor [5]. Hence, our discussion here will be in terms of phonemes.

II. Feature Set, Data Preparation, and Notation

Suppose we are interested in using the phonemes belonging to the set $\Omega = \{1, 2, \cdots, p\}$ uttered by the speakers belonging to the set $S_o = \{1, 2, \cdots, s\}$. We first choose a list of words which contain the phonemes in the set Ω. Each speaker in the set S_o utters every word in the list twice. The utterance is made in the ordinary laboratory environment with the computer running. The speakers were not coached in uttering the words. Every word utterance is recorded, sampled at the rate of 10 000 samples per second, and displayed on a RAMTEK display. Using the display, the word utterance is divided manually into nonoverlapping segments, neighboring segments corresponding to distinct phonemes.

Let N_{ij} stand for the total number of segments which correspond to the jth speaker and ith phoneme. We label these segments $g_{ij}^1, g_{ij}^2, \cdots$. The superscript in the label is assigned arbitrarily. Repeat this process for all the pairs $(i, j), i \in \Omega$ and speaker $j \in S_o$. Next, let us extract a vector of numerical features $\boldsymbol{\alpha}_{ij}^k$ from the segment g_{ij}^k. Typically, choose about N contiguous samples from the middle part of the segment which usually has samples numbering between 600 and 1500 and N is about 400. Label the samples as $y(1), y(2), \cdots, y(N)$. Assume that the time series $\{y(\cdot)\}$ obeys the following autoregressive model (1) where $w(\cdot)$ is a zero-mean white noise sequence with variance ρ, $w(t)$ being independent of $y(t-j)$ for all $j > 0$ and $\alpha_1, \cdots, \alpha_{n_1}$ are unknown parameters:

$$y(t) = \sum_{l=1}^{n_1} \alpha_l y(t-l) + (t). \tag{1}$$

$\hat{\alpha}$, the least squares estimates of the vector $\boldsymbol{\alpha}$ based on $y(1), \cdots, y(N)$, is [13]

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{t=n_1+1}^{N} z(t-1) z^T(t-1) \right]^{-1} \sum_{t=n_1+1}^{N} z(t-1) y(t)$$

$$z(t-1) = [y(t-1), \cdots, y(t-n_1)].$$

The required feature vector $\boldsymbol{\alpha}_{ij}^k$ is of dimension n, given below:

$$\boldsymbol{\alpha}_{ij}^k = \{\hat{\alpha}_1, \cdots, \hat{\alpha}_{n_1}, \hat{\alpha}_n\}\, T, \quad n = n_1 + 1, \hat{\alpha}_n = \hat{\rho}/s$$

$$\hat{\rho} = \frac{1}{N - n_1} \sum_{t=n_1+1}^{N} (y(t) - (\hat{\alpha})^T z(t-1))^2, s = \frac{1}{N} \sum_{t=1}^{N} y^2(t).$$

This procedure is repeated with all the segments g_{ij}^k for all possible values i, j, and k. We have chosen this particular feature set because it gave good results in practice. However, we stress that our decision rules are valid whatever may be the choice of feature vectors. In our case, $n_1 = 8, n = 9$.

We will consider three speech data bases labeled I, II, and III whose descriptions are given in Table I. The phonemes in the sets $\Omega^1, \Omega^2, \Omega^3$ are given in the Appendix.

The phonemes were chosen from words listed in the Appendix. The segmentation in data bases I and II was good. The segmentation in data base III was done by inspection by a student with an hour's training. The consequences of poor segmentation were twofold. The bad training samples lead to the poor estimates of the parameters in the decision rule. Since the test samples are not always what they claim to be, the actual error is much less than the value quoted in the tables.

TABLE I
Characteristics of the Three Speech Data Bases
$\Omega^1 = \{/i/, /l/, /\epsilon/, /æ/, /u/, /m/, /n/, /s/, /t/, /b/\}$
$\Omega^2 = \{/w/, /\ell/, /r/, /y/, /i/, /u/\}$, $\Omega^3 = \Omega^1 \cup \{/h/\}$

#	Number of speakers s	Ω	p	Average N_{ij}	Quality of Segmentation
I	4	Ω^1	10	40	Good
II	4	Ω^2	16	40	Good
III	6	Ω^3	11	260	Bad

III. Speaker Verification and Identification Methods

We will first consider speaker verification. As mentioned earlier, the speaker verification problem consists of determining whether the given (semantically unknown) utterance comes from a member of a known group of speaker S_o having s members in it or from an outsider. Besides the test utterance, the only available information about the group of speakers is their typical utterances of certain known words or phrases. Since by definition we cannot get any information about the outsider or imposter, we cannot use the Bayesian methods which require the probabilistic description of all the classes involved. We have to use the method of classical hypothesis testing. We will give two methods for speaker verification based on only one segment of the test utterance. Next we will generalize these methods to yield a decision based on all the phoneme segments in the utterance.

A. Speaker Verification Based on One Segment of an (Word) Utterance

From the waveform of unknown test utterance (word), we pick a segment having N (about 400) samples in it. Using these N samples, we will determine the corresponding feature vector $\boldsymbol{\alpha}$. We know only that the phoneme in that segment belongs to the set $\Omega = \{1, 2, \cdots, \rho\}$. We do not know anything about the speaker. We have to determine whether the speaker belongs to the set $S_o = \{1, 2, \cdots, s\}$ or not. If it does belong to S_o, we have to identify the particular member in it. The two methods to be presented are based on two different assumptions. We will consider some other possible assumptions and the corresponding decision rules in a later section.

Method 1: We assume that if $\boldsymbol{\alpha} \in (i, j)$ for some $i \in \Omega$ and $j \in S_o$ (i.e., if $\boldsymbol{\alpha}$ is from the phoneme i and speaker j), then $\boldsymbol{\alpha}$ obeys the following normal distribution:

$$\boldsymbol{\alpha} \sim N(\boldsymbol{\theta}_{ij}, \boldsymbol{S}_{ij}).$$

The n-vector $\boldsymbol{\theta}_{ij}$ and $n \times n$ matrix $\boldsymbol{S}_{ij}$ can be assumed to be known. If not, they can be estimated from the training samples as shown below. We will proceed with the test in two steps. In step 1, we determine the most likely phoneme-speaker pair in the set $\Lambda = \{(i, j): i \in \Omega, j \in S_o\}$ which could

have yielded the given α. This step is done using the likelihood rule. Let L_{ij} be the log likelihood associated with the phoneme i and speaker j. The most likely phoneme-speaker pair is denoted by (i^*, j^*).

$$(i^*, j^*) = \arg\left[\max_{(i,j)\in\Lambda} L_{ij}\right]. \tag{2}$$

In step 2, we determine whether the given segment could have come from an utterance belonging to (i^*, j^*). Here we have to use the classical hypothesis testing theory. Let

$$x_{ij} = (\alpha - \theta_{ij})^T S_{ij}^{-1} (\alpha - \theta_{ij}).$$

Then if the given segment comes from the phoneme-speaker pair (i^*, j^*), $x_{i^*j^*}$ obeys the following chi-squared distribution:

$$x_{i^*j^*} \sim \chi^2(n).$$

We choose arbitrarily a significance level β, say 0.95, i.e., we are fixing the probability of a given $\alpha \in (i, j)$ being misclassified to be no greater than $(1 - \beta)$. Then the threshold C_β can be obtained from the tables and the decision rule reads as follows:

$$x_{i^*j^*} \leqslant C_\beta \Longrightarrow \alpha \text{ belongs to } (i^*, j^*) \tag{3}$$

$$> C_\beta \Longrightarrow \alpha \text{ does not belong to } (i^*, j^*). \tag{4}$$

If (3) is true, we conclude that the given utterance belongs to an outsider.

Comment 1: The contingency in (4) can occur even if the speaker belongs to the set S_o, but the phoneme in α does not belong to Ω. Since we have ruled out the latter possibility, we are justified in making the above statement.

Comment 2: Geometrical interpretation. Let us draw the probability ellipses corresponding to the ps distributions for (i, j), $i \in \Omega$, $j \in S_o$ for the probability level β. If (3) is valid, then the point α falls within one of the ellipses. Obviously, the smaller the value of x_{i^*,j^*}, the greater is the credibility in the decision that the speaker is not an imposter.

Comment 3: The choice of the significance level β is highly controversial. As mentioned earlier, by choosing β, we are controlling only one type of error (error 1), namely, the error in classifying an utterance of a member of S_o as that of an imposter. There is another type of error (error II), namely, the error in classifying the utterance of an imposter as that of a member of S_o. Usually, the larger the β, the smaller is the probability of error I and the larger is the probability of error II. Good estimates of the probability of type II error depend on the characteristics of the imposter population employed in testing.

Comment 4: The decision rule used in step 2 is not the decision obtained by the Neyman-Pearson rule for classification among the ps classes (there are ps different possible pairs (i, j), $j \in \Omega$, $j \in S_o$). One may intuitively feel that if we employ a statistic in step 2 which is optimal in the Neyman-Pearson sense for the ps-class problem, we may get better results. This is an instance when our intuition is faulty [7]. As an illustration, consider the case of $p = 1$, $s = 2$, i.e., we want to see whether or not a given phoneme is from one of two speakers. Then the Neyman-Pearson decision rule for distinguishing between the two speakers is

$$x \triangleq \alpha^T S^{-1} (\mu_1 - \mu_2) \leqslant C_1 \Longrightarrow \alpha \in \text{speaker 1}$$

$$> C \Longrightarrow \alpha \in \text{speaker 2}.$$

The statistic x has different probability distributions depending on whether α belongs to speaker 1 or 2.

$$\text{If } \alpha \in \text{speaker 1}, x_1 = \frac{(x - \mu_1^T S^{-1} (\mu_2 - \mu_1))^2}{\| \mu_2 - \mu_1 \|_{S^{-1}}^2} \sim \chi^2(1)$$

$$\text{If } \alpha \in \text{speaker 2}, x_2 = \frac{(x - \mu_2^T S^{-1} (\mu_2 - \mu_1))^2}{\| \mu_2 - \mu_1 \|_{S^{-1}}^2} \sim \chi^2(1).$$

Rao [7] has shown in an example how a certain α which does not belong to either class can give a very small value for x_2 (of the order of 0.03) giving the erroneous decision that it belongs to speaker 2.

Comment 5: Estimation of θ_{ij} and S_{ij} from the training samples α_{ij}^k, $k = 1, 2, \cdots, N_{ij}$ defined earlier. Denote the corresponding estimates by $\hat{\theta}_{ij}$ and $\hat{S}_{ij}$.

$$\hat{\theta}_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} \alpha_{ij}^k$$

$$S_{ij} = \frac{1}{N_{ij} - 1} \sum_{k=1}^{N_{ij}} (\alpha_{ij}^k - \hat{\theta}_{ij})(\alpha_{ij}^k - \hat{\theta}_{ij})^T.$$

If N_{ij} is not very large, then $\hat{S}_{ij}$ may be a very poor estimate of S_{ij}. In such cases, we assume $S_{ij} = S_j$ for all $i \in \Omega$. An estimate of S_j is estimated as follows:

$$\hat{S}_j = \frac{1}{N_0 - 1} \sum_{i=1}^{P} \sum_{t=1}^{N} (\alpha_{ij}^t - \hat{\theta}_{ij})(\alpha_{ij}^t - \hat{\theta}_{ij})^T, \qquad N_0 = \sum_{i=1}^{P} N_{ij}.$$

In the decision rule, we will replace θ_{ij} and S_{ij} by their estimates.

Method 2: We assume that if α belongs to an (i, j), $i \in \Omega$ and $j \in S_o$, then α obeys the following distribution:

$$\alpha \sim N\left(\sum_{i=1}^{p} \sum_{j=1}^{s} \mu_{ij}\, \theta_{ij}, \sum_{i-1}^{p} \sum_{j=1}^{s} \gamma_{ij} S_{ij}\right) \tag{5}$$

where $\mu_{ij} > 0$, $\gamma_{ij} > 0$, $\Sigma_i \Sigma_j \mu_{ij} = \Sigma_i \Sigma_j \gamma_{ij} = 1$; otherwise μ_{ij}, γ_{ij} are arbitrary. This assumption is less restrictive than the assumption in Method 1. θ_{ij} and S_{ij} can be estimated as in Method 1. Let us consider the particular case when

$$\gamma_{ij} = \frac{1}{ps}, S = \frac{1}{ps} \sum_{i=1}^{p} \sum_{j=1}^{s} S_{ij}$$

$$x = \min \left(\alpha - \sum_{ij} \mu_{ij}\, \theta_{ij}\right)^T S^{-1} \left(\alpha - \sum_{ij} \mu_{ij}\, \theta_{ij}\right).$$

If $ps = 2$, then $x \sim \chi^2(n - 1)$, according to [7], whatever may be the value of μ_{ij} in (5). As long as $n > ps$, it is reasonable to assume that x obeys the distribution $\chi^2(n - ps + 1)$. Using the test statistic x, we can design a threshold rule as before.

The difference between Methods 1 and 2 can be pictured geometrically as in Fig. 1. If the vector α falls between any two small ellipses in Fig. 1, it is excluded from S_o by Method

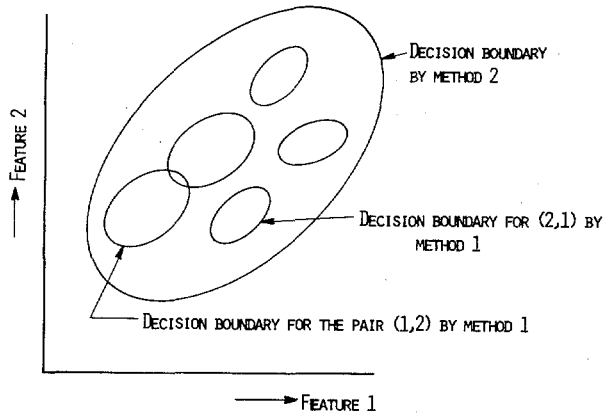

Fig. 1. The decision boundaries for speaker verification by Methods 1 and 2 in Section III-A.

1, whereas Method 2 places it in the class S_o as long as it is within the large ellipse.

B. Speaker Verification Rules Based on An Entire (Word) Utterances

The unknown test utterance (word) is divided into m segments and the segments are numbered as $1, 2, 3, \cdots, m$ from left to right. The segmentation need not be done as carefully as in Section II. The only condition is that we should be sure that there is no more than one phoneme in a segment. Let $\boldsymbol{\alpha}(k)$ denote the feature vector extracted from the kth segment. Our rule is based on the vectors $\{\boldsymbol{\alpha}(1), \cdots, \boldsymbol{\alpha}(m)\}$.

We assume that the phoneme in the segment k is known to belong to $\Omega_k \in \Omega$.

Rule 1, Majority Rule: We consider the m segments of the utterance individually. If a majority of the segments are classified as not belonging to S_o, then the entire utterance is declared as not belonging to S_o.

Rule 2, Weighted Majority Rule: This is also a majority rule except that the decisions obtained by the various segments are weighted. The weights are determined by the fact that not all phonemes are equally effective for speaker verification or identification. Similarly, if successive segments belong to the same phoneme, we have to reduce the weight for such segments. If the phoneme decision for the kth segment is $j_k^* \in \Omega$, then the speaker decision is made using the statistic x:

$$x = \frac{\sum_{k=1}^{m} w(i_k^*) d_k}{\sum_{k=1}^{m} w(i_k^*)}$$

where $w(j)$ is the weight attached to phoneme j, $0 \leqslant w(j) \leqslant 1$, and d_k is equal to 1 if the kth segment is classified as belonging to imposter, and is zero otherwise. Hence, if $x \geqslant \frac{1}{2}$, the utterance is declared as not belonging to S_o. The weights $w(j)$ have to be determined empirically.

Rule 3: Here we will not make decisions using the individual segments separately. Rather we will obtain one decision based on the entire utterance.

$$x_j = \sum_{k=1}^{m} \min_{i \in \Omega_k} (\boldsymbol{\alpha}(k) - \boldsymbol{\theta}_{ij})^T \boldsymbol{S}^{-1} (\boldsymbol{\alpha}(k) - \boldsymbol{\theta}_{ij})$$

$$x = \min_{j \in S_o} x_j, \; j^* = \arg \left[\min_{j \in S_o} x_j \right].$$

If x indeed comes from the speaker j^*, then $x \sim \chi^2(mn)$.

As before we can have the following decision rule at the significance level β:

$x \leqslant C_\beta \Longrightarrow$ utterance comes from speaker j^*

$> C_\beta \Longrightarrow$ utterance belongs to an imposter.

Discussion: If the decision is to be unambiguous, all three rules must yield the same decision. Rule 3 can be used only if we are sure that the phonemes in the various segments do belong to the respective sets $\Omega_1, \Omega_2, \cdots, \Omega_m$. If even one of them does not belong to its respective set, then the statistic x may become relatively large, leading to the decision of imposter even though it is not true. Rule 1 (or preferably Rule 2) is free from this drawback. When we are dealing with known utterances, Rule 3 is superior to Rules 1 or 2.

C. Tests for Speaker Identification

It is to be noted that the various rules given in the earlier section yield an explicit decision for the most likely speaker in the set S_o. One can give other methods based on the Bayesian viewpoint as well. Since this topic is fairly routine in the statistical literature, we will not repeat it here.

IV. Numerical Results

We will use the three data bases mentioned in Section II. Typically, we will use half the data for estimating the parameters in the decision rule and use the remaining half as test segments. We compare the decision given by the rule for the test segments with their known classification. We assume $\boldsymbol{S}_{ij} = \boldsymbol{S}_j$ for all $i = 1, \cdots, p$.

We first consider speaker identification with data base II. Using only one segment of the utterance and with the set $\Omega = \Omega^2$, the speaker identification accuracy is 95 percent. The corresponding decisions are summarized in Table II. Next we repeated the experiment with a part of data base 1 with $\Omega = \Omega^4 \triangleq \{/i/, /\varepsilon/, /æ/, /u/, /l/, /m/, /n/\}$ using only the data of the phonemes in Ω^4. The recognition accuracy is about 96 percent and the decision matrix is in Table III. Next we considered data base III with $\Omega = \Omega^3$. The recognition accuracy is 75 percent and the decision matrix is in Table IV. In Tables II-IV, Method 1 was employed.

If we use an entire utterance consisting of four phonemes, the identification error with data base III is less than 1 percent. The corresponding error rates with data base I using words was also considerably less than 1 percent.

Our results with speaker verification were performed by determining the decision rules using $(s - 1)$ speakers and testing the utterance of the sth speaker who is left out, using the decision rule. Typical error rates based on one phonemic segment of the utterance are comparable to the off-diagonal terms of Tables II-IV.

TABLE II
DECISION MATRIX FOR SPEAKER IDENTIFICATION WITH DATA BASE II

actual speaker \ decision	1	2	3	4
1	28	0	1	1
2	0	29	1	0
3	1	1	28	0
4	1	0	0	20

TABLE III
DECISION MATRIX FOR SPEAKER IDENTIFICATION WITH PART OF DATA BASE I AND $\Omega = \Omega^4$

actual speaker \ decision	1	2	3	4
1	62	1	2	1
2	4	63	0	2
3	2	2	58	8
4	1	0	1	68

TABLE IV
DECISION MATRIX FOR SPEAKER IDENTIFICATION WITH DATA BASE III

actual speaker \ decision	1	2	3	4	5	6
1	190	3	15	8	13	20
2	5	153	3	4	29	29
3	13	4	201	12	1	18
4	7	7	15	183	6	38
5	16	27	13	5	200	15
6	17	14	17	31	10	168

V. SPEAKER VERIFICATION AND IDENTIFICATION WITH A LARGE NUMBER OF SPEAKERS

When we are dealing with a large number of speakers, say a thousand or more, the methods described earlier are clearly impractical and need modification. One method is to subdivide the speaker set into a number of small groups in a meaningful way. We can first test whether the utterance belongs to any one of these groups using the techniques described earlier. If we do find that the utterance does belong to a subset, we can determine the identity of the speaker by the techniques described earlier. The problem of grouping the speakers is commonly labeled as clustering and there are a number of algorithms for accomplishing it. The direct use of these techniques using the given data bases is futile because of the considerable variation in the data due to the variety of phonemes which are not of direct use to us. In clustering for our problem, we should ignore the variation of the data due to phonemes and concentrate on the interspeaker variation in the data so that "similar" speakers will be grouped together. To do this, we need to extract a vector $\boldsymbol{\alpha}_j$ from each pattern vector $\boldsymbol{\alpha}_{ij}$, so that $\boldsymbol{\alpha}_j$ depends only on the speaker j and not on the phoneme i. Thus, we are led to the study of the phoneme-speaker variation in speech.

The hypothesis used in the earlier section can be restated as (6):

$$\boldsymbol{\alpha}_{ij}^k = \boldsymbol{\theta}_{ij} + \boldsymbol{\eta}_{ij}^k \tag{6}$$

where $\{\boldsymbol{\eta}_{ij}^k,\ k = 1, 2, 3, \cdots\}$ is a sequence of independently identically distributed (i.i.d.) vectors with zero mean and covariance matrix $\mathcal{S}_{ij}$.

Let us consider some alternative hypotheses in which vector $\boldsymbol{\theta}_{ij}$ can be expressed as an explicit function of the contributions of the speaker j and phoneme i. Among them, the simplest hypothesis is the additive hypothesis in (7) where $\boldsymbol{\theta}_i$, $\boldsymbol{\alpha}_j$ are n-vectors depending on i and j, respectively:

$$\boldsymbol{\alpha}_{ij}^k = \boldsymbol{\theta}_i + \boldsymbol{\alpha}_j + \boldsymbol{\eta}_{ij}^k. \tag{7}$$

Another of the seperative hypotheses is the multiplicative hypothesis (8):

$$\boldsymbol{H}_j \boldsymbol{\alpha}_{ij}^k = \boldsymbol{\beta}_i + \mathcal{E}_{ij}^k \tag{8}$$

where $\boldsymbol{H}_j$ is an $n_1 \times n$ matrix, $\boldsymbol{\theta}_i$ is an n_1-vector, and $\{\mathcal{E}_{ij}^k\}$ is an n_1-dimensional zero mean i.i.d. sequence. In (8) we regard the speaker as providing a transformation matrix which rotates the feature vector due to the phoneme only to produce the observed feature vector $\boldsymbol{\alpha}$. The dimension n_1 can be chosen by computational convenience.

Still another hypothesis is the additive-multiplicative hypothesis in (9):

$$\boldsymbol{\alpha}_{ij}^k = \boldsymbol{\theta}_i + \gamma_i \boldsymbol{\alpha}_j + \boldsymbol{\eta}_{ij}^k \tag{9}$$

where γ_i is a scalar depending only on i.

The hypotheses in (6)-(9) can be compared using the standard theory of hypothesis testing and the data based I-III mentioned earlier. This has been done by the author in [8] and the conclusion is that hypothesis I has to be preferred to others at the usual 95 percent significance level. This is not entirely unexpected.

Our next problem is to see which of the seperative hypotheses is the "best" approximation for our problem. This can be done by designing optimal speaker recognizers based on the hypothesis in (7)-(9) and comparing their recognition capabilities with that based on the hypothesis in (6). The details of the construction of the optimal recognizers are given in [8]. The result is that the additive hypothesis in (7) approximates the model in (6) better than other hypotheses. In Tables V and VI, we have the decision matrices for the speaker identification using the recognizers based on the hypotheses (7) and (9) using the part of data base I used in Section IV with $\Omega = \Omega^4$. Comparison of Tables III, V, and VI shows that every diagonal element of the matrix in Table III is always (and often significantly) greater than the corresponding elements in the matrices of Tables V and VI. The differences between Tables V and VI are not significant. Hence, we prefer the hypothesis in (7) to the hypothesis in (9) because the former is computationally easy.

When we have a large number of speakers, we can obtain the $\boldsymbol{\alpha}_j$, $j = 1, 2, \cdots s$ for all the speakers using the hypothesis in (7) as discussed above, and cluster these vectors and determine the speakers who constitute the members of the various subgroups.

TABLE V
DECISION MATRIX FOR SPEAKER IDENTIFICATION WITH DATA BASE I BASED ON HYPOTHESIS (7)

actual speaker \ decision	1	2	3	4
1	55	4	5	2
2	2	63	4	0
3	2	12	47	1
4	1	3	9	57

TABLE VI
DECISION MATRIX FOR SPEAKER IDENTIFICATION USING A PART OF DATA NASE I USING HYPOTHESIS (9)

actual speaker \ decision	1	2	3	4
1	56	3	6	1
2	4	61	4	0
3	1	8	52	1
4	3	4	2	51

VI. THE CHOICE OF PHONEME FOR SPEAKER RECOGNITION

In the literature, one finds suggestions that some phonemes, like nasals or some vowels, are the best phonemes for speaker recognition [11]. We will investigate these claims. There are two distinct questions. The first question is whether all the information needed for speaker identification or verification is contained in special subclasses of phonemes like nasals. The numerical results quoted in the earlier section clearly indicate that the answer to this question is a clear no. All the phonemes contain information useful for speaker recognition. The second question is the determination of the phonemes which may carry more information than others for speaker identification. This question can be investigated in two ways. First of all, we could perform two different recognition experiments of the type discussed in Section III with the corresponding phoneme sets Ω labeled Ω' and Ω''. Let Ω'' have all the phonemes in Ω' and an additional phoneme. The comparison of the correct recognition rates in the experiment clearly indicates whether the additional phoneme in Ω'' is relatively useful for speaker recognition in comparison with the phonemes of the set Ω'. For instance, let us perform the two experiments with data base I, with $\Omega' = \Omega^4$ and $\Omega'' = \Omega^4 \, U \{/s/, /t/, /b/\}$. The recognition error rates with Ω' and Ω'' are, respectively, 4 and 10.6 percent. The corresponding decision matrices are in Tables III and VII, respectively. This clearly indicates that the phonemes /s/, /t/, and /b/ are less useful for speaker recognition than the vowels and nasals in Ω^4.

The second method is to compare the different phoneme-speaker pairs by a suitable distance function which measures the discrepancy between two phonemes spoken by two different speakers. For instance, suppose we want to compare the effectiveness of phoneme i to distinguish between two speakers j_i and j_2. Then the distance function can be

$$D_i(j_1, j_2) = \| \boldsymbol{\theta}_{ij_1} - \boldsymbol{\theta}_{ij_2} \|^2_{S_i^{-1}}$$

TABLE VII
DECISION MATRIX FOR SPEAKER RECOGNITION USING DATA BASE I WITH $\Omega = \Omega'' = \Omega^4 \, U \{/s/, /t/, /b/\}$

actual \ decision	1	2	3	4
1	98	2	3	3
2	5	87	3	2
3	4	3	88	3
4	5	2	2	104

where

$$\boldsymbol{S}_i = (\boldsymbol{S}_{ij_1} + \boldsymbol{S}_{ij_2})/2.$$

Ideally we should prefer phoneme i_1 to phoneme i_2 if (10) is valid.

$$(D_{i_1})_{jk} > (D_{i_2})_{jk}, \quad j = 1, \cdots, s, k = j+1, \cdots, s. \tag{10}$$

There may exist a pair of phonemes which satisfy (10), but there does not exist phoneme i_1 which obeys (10) for all i_2. Hence, we try to construct a scalar index, say a_i, for each phoneme i using the matrix D_i so that a_i is greater than a_j means that (10) is valid for most j, k. Two choices are a_i, b_i:

$$a_i = \min_{j,k} (D_i)_{jk}, \quad j, k \in S_o, j \neq k$$

$$b_i = \sum_{j,k} (D_i)_{jk}.$$

In a given set of phonemes, we should choose that phoneme for speaker recognition which has the greatest values for both the indices a and b. As an illustration, let us consider the vowels and nasals in data base I. The indices a and b for all these phonemes are given in Table VIII.

From Table VIII, the phoneme /i/ is the "worst" of the lot and the phonemes /l/ and /ɛ/ are the "best" of the lot. The phoneme /l/ is superior to /i/ even by the criterion (10). Studies by the first method have confirmed that /l/ and /ɛ/ are more useful than /i/ for speaker recognition. We also see from Table VIII that the nasals /m/, /n/ are not necessarily superior to the vowels /l/ and /ɛ/ for speaker recognition.

VII. PHONEME RECOGNITION WITH UNKNOWN SPEAKERS

Currently there are a number of research projects [9], [10], [12] which deal with the recognition of words or phrases from an utterance. Many of these studies deal with utterances from two or three speakers, and the methods of recognition do not incorporate the variability in the waveforms of the same phoneme spoken by different speakers. As such, these methods may not be able to handle utterances from a large number of speakers. Out intention is to show [12] that the accuracy of recognition of certain phonemes can be effectively increased by recognizing the variety present in the utterances of the same phoneme by different speakers.

Let us first design the phoneme recognizer ignoring the variation in the speech due to speaker. In that case, the model of speech is:

$$\boldsymbol{\alpha}_{ij}^k = \boldsymbol{\theta}_i + \boldsymbol{\eta}_{ij}^k \tag{11}$$

TABLE VIII

phoneme	/i/	/ɪ/	/ɛ/	/æ/	/u/	/m/	/n/
index a	250	1510	716	169	231	662	527
index b	5423	20021	23662	16832	9790	14594	9158

where $\{\eta_{ij}^k,\ k = 1, 2, \cdots\}$ is an i.i.d. sequence with normal distribution and $N(O, S_{ij})$ and θ_i depend on only the phoneme index i and not on the speaker index j. θ_i can be easily estimated from the training sample $\{\alpha_{ij}^k, k = 1, \cdots n\}$:

$$\hat{\theta}_i = \frac{1}{Ns} \sum_{k=1}^{N} \sum_{j=1}^{s} \alpha_{ij}^k, \quad i \in \Omega.$$

Let L_i = log likelihood of the test utterance with pattern vector α assumed to belong to phoneme i. Then the phoneme decision $\bar{i}$ obtained for the test utterance is obtained as follows:

$$\bar{i} = \arg\,[\max_i L_i]. \tag{12}$$

We should statistically compare the decision $\bar{i}$ with the decision i^* obtained in (2) utilizing the speaker information as well. This comparison is done using data base I. In Tables IX and X we give the results of decisions i^* and $\bar{i}$, respectively.

Tables IX and X reveal the superiority of the rule in (2) since every diagonal element in Table IX is strictly greater than the corresponding element in Table X. With phonemes such as /ɛ/ or /l/ the discrepancy between them could be as much as 30 percent.

From Tables IX and X we can also notice the close relationship between the role of sensitivity of phoneme recognition to speaker variation and the choice of phoneme for speaker recognition, as discussed in Section VI. If a phoneme is not relatively good for speaker recognition, it means that the variation in the utterances of the same phoneme due to different speakers is not substantial. In such cases the decisions i^* and $\bar{i}$ should not be too different. This is the case with a phoneme /i/. On the other hand, if a phoneme is very good for speaker recognition, then the variability in its utterances due to speaker is large and consequently phoneme decision $\bar{i}$, which ignores such variation, is inferior to the decision i^* which incorporates the speaker variation concept in it. For instance consider the phonemes /l/ and /ɛ/ which were found to be relatively good for speaker recognition. Hence, the error with decision $\bar{i}$ should be greater than the error with decision i^* for these two phonemes. The truth of this statement is demonstrated in Tables IX and X.

VIII. Conclusion

We have shown conclusively that the use of statistical decision theory leads to relatively simple decision rules for both speaker verification and identification. We can use either one phonemic segment or the entire utterance. We have considered in some detail the speaker-speech interaction. The results of such a study are very useful for a variety of purposes such as improving the accuracy of speech recognizers by allowing for speaker variability, methods of clustering speakers for handling speaker verification problems with large groups of persons, etc.

TABLE IX
Phoneme Recognition Using Rule (2)

actual phoneme \ Decision →	/i/	/ɪ/	/ɛ/	/æ/	/u/	/m/	/n/
/i/	33	0	0	0	0	0	0
/ɪ/	0	28	4	0	0	0	0
/ɛ/	0	0	27	0	0	0	0
/æ/	0	0	7	20	0	0	0
/u/	0	0	0	0	5	0	2
/m/	0	0	0	0	1	29	7
/n/	0	0	1	0	1	3	53

TABLE X
Phoneme Recognition Using Rule (12)

actual phoneme \ Decision	/i/	/ɪ/	/ɛ/	/æ/	/u/	/m/	/n/
/i/	30	3	0	0	0	0	0
/ɪ/	1	21	6	4	0	0	0
/ɛ/	0	4	16	6	0	0	1
/æ/	0	2	6	19	0	0	0
/u/	0	0	4	0	41	2	6
/m/	0	0	0	0	1	27	9
/n/	1	0	0	1	3	10	43

We have assumed that an utterance is manually divided into phonemic segments without necessarily having to know the phoneme in each segment. Present studies are aimed at removing this restriction.

Appendix

List of Phonemes Used in This Study

Phoneme	Word having the phoneme
/i/	mean
/I/	mint
/ɛ/	men
/æ/	mat
/u/	moon
/w/	away
/y/	buyer
/ɻ/	alas
/r/	direct
/m/	man
/n/	nest
/b/	boon
/s/	soon
/t/	team
/h/	his

List of Words Used in Data Bases I and III

men, mean, moons, mats, meant, mint, meat, mess, moon, gnat, noon nest, bit, bean, boon, boob, bet, boot, bamboo, bent, beast, boost, behest, team, tune, tin, tent, test, snoot, sits, soon, seem, am, last, an, hit, him, hoot, hint, habit.

Acknowledgment

The author is indebted to Dr. M. Mittal and P. S. Ramakrishna for assistance with numerical computations.

References

[1] P. D. Bricker *et al.*, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, 1970.
[2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, p. 1304, June 1974.
[3] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 169–175, Apr. 1975.
[4] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, Apr. 1975.
[5] R. L. Kashyap *et al.*, "Spoken work recognition system of Purdue," Advanced Automation Res. Lab., School of Elec. Eng., Purdue Univ., Lafayette, IN, AAR Memo 15, Oct. 1975.
[6] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.
[7] C. R. Rao, *Linear Statistical Inference*. New York: Wiley, 1965.
[8] R. L. Kashyap, "The separation of phonemic and speaker components of speech," unpublished, 1976.
[9] R. W. Becker and F. Poza, "Acoustic phonetic research in speech understanding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 416–426, Oct. 1975.
[10] C. C. Tappert and N. R. Dixon, "A procedure for adaptive control of the interaction between acoustic classification and linguistic decoding in automatic recognition of continuous speech," *Artificial Intelligence*, vol. 5, pp. 95–113, 1974.
[11] J. W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Amer.*, vol. 43, pp. 368–372, Feb. 1968.
[12] R. L. Kashyap and M. C. Mittal, "Word recognition in a multitalker environment using syntactic methods," in *Proc. 3rd Int. Conf. Pattern Recognition*, San Diego, CA, 1976.
[13] R. L. Kashyap and A. R. Rao, *Dynamic Stochastic Models From Empirical Data*. New York: Academic, 1976.

LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise

MARVIN R. SAMBUR AND NUGGEHALLY S. JAYANT, MEMBER, IEEE

Abstract—**An important problem in some communication systems is the performance of linear prediction (LPC) analysis with speech inputs that have been corrupted by (signal-correlated) quantization distortion or additive white noise. To gain a first insight into this problem, a high-quality speech sample was deliberately degraded by using various degrees (bit rates of 16 kbps and more) of differential PCM (DPCM), and delta modulation (DM) quantization, and by the introduction of additive white noise. The resulting speech samples were then analyzed to obtain the LPC control signals: pitch, gain, and the linear prediction coefficients. These control parameters were then compared to the parameters measured in the original, high quality signal. The measurements of pitch perturbations were assessed on the basis of how many points exceeded an appropriate difference limen. A distance measure proposed by Itakura was used to compare the original LPC coefficients with the coefficients measured from the degraded speech. In addition, the measured control signals were used to synthesize speech for perceptual evaluation. Results suggest that LPC analysis/synthesis is fairly immune to the degradation of DPCM quantization. The effects of DM quantization are more severe and the effects of additive white noise are the most serious.**

Manuscript received December 29, 1975; revised March 10, 1976 and July 1, 1976.

The authors are with Bell Laboratories, Murray Hill, NJ 07974.

I. Introduction

With the increased reliance on the digital communication of speech, it has become important to investigate the tandem operation of different digital coding systems. One of the more interesting tandem applications is the linking of a linear prediction (LPC) vocoder system [1] with differential speech quantizers [2]. For example, military communication systems frequently require, because of different bandwidth constraints in different parts of the system, the linking of a 16 kbps differential pulse-code modulation (DPCM) or delta modulation (DM) coder with an even more efficient 2.4 kbps linear prediction vocoder. It is the purpose of this paper to examine the compatibility of typical waveform coders and linear prediction vocoder systems. The major portion of this paper will be devoted to the study of a tandem connection going from waveform coder to LPC vocoder. This examination will be concerned with the performance of a LPC vocoder with speech inputs derived from various degrees (between 16 and 40 kbps) of DPCM or DM quantization of a high quality speech sample. To add generality to our investigation, we have