NEW TECHNIQUES FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Larry L. Pfeifer

Signal Technology, Inc.
15 W. De La Guerra Street
Santa Barbara, California 93101

Abstract

A method for text-independent speaker identification has been developed which utilizes vowel sounds as the basis for extracting speaker characteristics. Using 63 minutes of conversational speech data from 20 speakers, it was found that vowel recognition is not necessary. Instead, the vowel samples can be pooled such that they represent each person's vowel space, which is expected to be very speaker-dependent. A sequential analysis process has improved the decision procedure by allowing vowel samples to be tested until a specified level of confidence is reached in the identification. This dynamic decision procedure is similar to a human perception process where we can quickly identify a unique voice, but listen longer when there is uncertainty.

Introduction

The use of vowel sounds for speaker identification typically requires that vowel samples first be classified according to vowel category, so that vowels of the same category can be compared in the speaker identification process. It has been demonstrated, however, that it is only necessary to detect vowel-like sounds in the speech material and that speaker identification performance actually improves when there is no vowel recognition. The elimination of vowel recognition from this approach to speaker identification reduces the procedural complexity of the process, and removes an extra decision stage which would have compounded the overall system error rate. Some explanation for the improved identification performance is that the pooled vowel samples from each speaker are a representation of each person's vowel space, plus effects of vowel categorization errors are eliminated.

Another significant outcome of this research was the successful application of sequential analysis to the decision process. Sequential analysis relies on the accumulation of speaker classification results from several vowel samples before making a decision. If a sufficient number of test samples are classified as any one speaker, then a decision can be made, with a certain level of confidence, regarding the identity of the unknown speaker. The method allows acceptance and rejection thresholds to be established with specified error probabilities. This makes the sequential analysis procedure a dynamic process which accumulates and tests vowel samples until a confident decision can be made. Thus, the decision time and the amount of speech material needed is variable, depending on the speaker being tested. A speaker whose voice is distinct would be identified in a short period of time, whereas one whose voice is similar to others might require a larger number of samples, and therefore more time.

Methods and Procedures

The experiments which led to the conclusions of the research were based upon a very large data base of vowel samples from more than one hour of speech material excerpted from recorded interviews. Recordings were made of twenty speakers, ten male and ten female. Two of the speakers were recorded two-weeks later so that some effects of time could be studied as well. The interview sessions were informal, with the interviewer asking some general questions and the subject responding in a fashion that usually lead to a conversational situation. The data is a representative sampling of casual speech that would be found in natural language communication. There was no control over the content of the speech material other than the general semantic direction determined by the question.

While each interview lasted at least 12 minutes, a 3-minute segment from the initial portion of each was used as a source for vowel samples. A 1.5-minute

---

This work was performed while the author was at the Speech Communications Research Laboratory, Inc., 800A Miramonte Drive, Santa Barbara, California 93109.

This work was sponsored by Rome Air Development Center, Air Force Systems Command, Griffiss Air Force Base, New York under contract F30602-76-C-0157.

CH1285-6/78/0000-0283$00.75@1978IEEE

segment was used from the two repeated interviews. These 22 segments provided a total of 63 minutes of recorded conversation, consisting mostly of subject speech, plus some instances of extended pauses, laughter, and interviewer questions or comments. Each of the 22 segments was digitized at a sampling frequency of 10 kHz and stored in computer files. Using the Interactive Laboratory System (ILS) described by Pfeifer (1977), the speech files were scanned for occurrences of the vowels /i,ɪ,ɛ,æ,ə/, which were then labelled with vowel identity, stress, environment, sequence number, word in which the vowel occurred, and speaker initials. These particular vowel sounds were chosen because they are among the most frequently occurring vowels in spoken English and therefore provided a maximum of vowel samples for a given amount of speech.

The collection of vowel samples was augmented by two automatic procedures which significantly reduced the variability and inconsistencies in vowel labelling. The first procedure was an automatic boundary detector. This is essentially a segmentation algorithm which was beneficial in helping the operator locate the desired vowel sounds. While the operator still had to identify the vowels, the algorithm provided a consistent criterion for locating vowel boundaries, based upon the location of maximum spectral change in the speech signal.

The second procedure which reduced variability in vowel labelling was an automatic algorithm for locating the most stable portion of a vowel sound. This is an important step because it results in a consistent criterion for where to perform the analysis within the vowel. The steady-state is designated as that location within a vowel segment where there is minimal spectral change in the speech signal.

The five-vowel data base consists of 4786 labelled vowel samples. These data were split into two independent sets of reference and test samples. The autocorrelation method of linear prediction was used to analyze a 20 msec window in the steady-state portion of each vowel. Twelve reflection coefficients (k-parameters) were generated and used as feature vectors.

The distance metric used was the weighted Euclidean distance. With this form of distance measure, the reference data for each speaker is represented by a mean vector and an inverse covariance matrix. In a closed-choice speaker identification task, where the test sample is assumed to be a member of the set of reference set, a minimum distance criterion is sufficient for the classification process.

Results

Experiments Assuming Vowel Recognition.

Each of the 20 speakers was represented by five references, one for each of the five vowels. If a test vowel from an unknown speaker could first be categorized as to vowel class, then it would be matched against the 20 references (one for each speaker) for the corresponding vowel class. The results of the individual vowel category tests were then grouped to provide overall 5-vowel results based upon the assumption of correct vowel recognition. There was a total of 2221 test samples from all five vowel categories. On the basis of classifying each vowel sample according to speaker, the overall score was simply computed from the scores of the separate vowel experiments, i.e., 871 correct out of 2221 (39.2 percent). A composite confusion matrix for all five vowels is shown in Figure 1.

By traditional scoring, 39.2 percent correct classification might be considered unacceptably low. However, it is not the classification of any one input sample that is significant, for it may not be wise to make an identification decision based upon a single vowel sample. Instead, it is better to examine the accumulated results from the classification of many vowel samples and then make an identification decision based upon the distribution of the individual classifications. This is an acceptable concept for it means that the decision would be global to a particular amount of speech material. Thus, the test data from speaker HAN in Figure 1 might consist of 130 vowels from 90 seconds of speech, and based upon a simple majority decision or the mode of the classification counts, it could be concluded that the 90 seconds of speech belongs to speaker HAN. The mode of each row of the matrix has been outlined in Figure 1, and it can be seen that a modal decision rule would make only one error out of 20 decisions. With this technique, the strength of a decision is a function of the largest classification count and the next largest classification count.

Experiments Without Vowel Recognition.

The task of performing vowel recognition prior to speaker identification was eliminated by representing each speaker with a single reference pattern consisting of the pooled samples of all five vowel categories, and then testing with unclassified vowel samples.

An experiment using this approach showed that of the 2221 test samples, 999 were classified as the correct speaker (44.98 percent). This result is an improvement over those requiring total

vowel recognition, and it is statistically sound because of the pooling of the vowel samples. It can be seen in the confusion matrix of Figure 2 that the mode of the classifications is located at the true speaker in all cases except one. This is the same speaker who was missed when vowel recognition was employed. In terms of using a modal decision rule to actually make the speaker identification decisions, the results are essentially unchanged with or without vowel recognition. But since the number of correct classifications increased when there was no vowel recognition, it is expected that the mode of the classifications for each test speaker would also increase in strength. A comparison of Figure 1 and Figure 2 shows that with no vowel recognition the strength of the mode increased for 15 test speakers,

	JOE	RHF	HAN	MOM	DJB	MAE	EHH	MBB	LLP	JDM	JAC	BFH	HS	BTO	NAT	ECJ	CMW	SAD	JME	JC
JOE	22	1	6	4		2	2	11	13	1	1		3	3	3	7	2	4	1	4
RHF	1	73	10	5	14	25	11	5	4	19		1	3	1	3	13		1		
HAN	1	2	71	2	9	6	4	2	15	12		1	1		2	2				
MOM	7	8	4	41	11	9	3	6	7	6	1		3	1		8		2		3
DJB		1	13	2	33	19	5	2	6	10		2	1		4	6	1	1		3
MAE		5	9	4	9	32	4	7	16	7		4	1		3	8				3
EHH	1	7	9	3	9	15	19	9	6	5	1	1			1	10		1		1
MBB	2	5	1	9	5	30	5	20	5	2		1			5	14		3	2	1
LLP	2	1	20	3	7	12			36	5		1	3	5	4	5	2			5
JDM		3	11	5	4	4	3	2	3	85		1	1			1		3	1	
JAC		3	1	7	1		1	3	1		72	2	1	1	2	8	1	3		4
BFH		1	3		2	3	1	1	9	3	1	65	15		11	14		1	1	11
HS	1	4	4		2	2			1	5		18	82	3	5	5		5	1	6
BTO	3	1	1			1	2	2	5	1		4	3	28	21	9	3	3		11
NAT		2	2			6			5	1		3	6	12	26	7	7		1	13
ECJ	1	4	4		2	13	2	1	3	2	1	6	4	2	4	62	1			2
CMW	1		1		1	4			2	2		1		5	16	6	25	6	2	3
SAD	1		1						1	1		1	2	4	9	3	8	21	3	4
JME		2				3			2	3	1	10	2	2	8	4	5	11	24	2
JC	8	2	2	1	5	12	2	3	1	2		6	2	1	7	11	4	2		34

Figure 1. Confusion matrix for text-independent speaker identification, assuming vowel recognition. There are five references for each speaker, one corresponding to each vowel class. Each input test sample is first classified according to vowel category and then matched with those references for that vowel only.
CLASSIFICATION SCORE:
871 out of 2221 correct (39.2%)
MODAL SCORE:
19 out of 20 correct (95%)

	JOE	RHF	HAN	MOM	DJB	MAE	EHH	MBB	LLP	JDM	JAC	BFH	HS	BTO	NAT	ECJ	CMW	SAD	JME	JC
JOE	31		1	2	4	3		1	20	1	2			2	2		7	6		8
RHF		43	5	4	23	37	14	7	4	15			2		2	27	2		2	2
HAN	2		46	4	24	10	2		8	26		1				4		4		
MOM	6	3		53	4	9	1	8	8	13	3			1		6	3	1	1	
DJB			5	1	61	16		2	12	3			2	1	1	2				3
MAE		1	5		21	47		5	15	6		1	1			5	1		1	3
EHH	2	2	9	4	19	15	24	1	4	4				1		10	1	1		1
MBB	4	1	2	3	7	32	3	24	4	2	1				2	18		1		6
LLP	2		7	1	8	18			61	1				3	4	3	1			2
JDM			1	2	8	4	5		1	100						4		1		1
JAC	2	3			1	1		2			73			2		13	5	6	3	
BFH					4	2			13			64	7	4	12	23	2	1	4	9
HS		2	3		3	2	2	1	1	1		9	66	6	5	24		12		7
BTO				1					4		1		2	46	16	4	11	3	1	9
NAT						2		1	9			2	6	16	26	3	11	5		11
ECJ		1			2	8	3	2	1			3	2	1	5	84	3			1
CMW														1	7	4	48	12	1	2
SAD	1					1								3		1	17	35	1	
JME							2	8	1		2	1			1	5	9	19	28	3
JC	5		1		1	8		4	5			1	1	1	24	7	3	5		39

Figure 2. Confusion matrix for text-independent speaker identification assuming no vowel recognition. There are 20 reference patterns, one for each speaker. Each reference is made up of the pooled samples of all five vowel classes.
CLASSIFICATION SCORE:
999 out of 2221 correct (44.98%)
MODAL SCORE:
19 out of 20 correct (95%)

decreased for four speakers, and remained unchanged for one speaker.

Sequential Analysis.

The simulation of a true text-independent speaker identification situation with conversational speech requires that, for the preparation of speaker references, vowel samples be collected for a certain period of time, and that the test vowel samples be taken from the unknown recording in a sequential fashion as they occur in the speech. As each test sample is classified, the cumulative results of all the classifications can be examined to determine if a sufficient number of samples have been placed with any one reference speaker. This type of sequential analysis procedure was described by Wald (1952) and it is appropriate to this application where there are many samples from an unknown speaker. This is also known as a delayed decision procedure because the actual decision is not made until a sufficient number of individual classifications have been performed.

In order to apply the sequential analysis procedure, we must examine the accumulated intermediate speaker classification results after each test sample has been classified. Thus, the intermediate results reflect the status of the decision process as a function of time. For a given set of test samples from an unknown talker, each test sample is subjected to a minimum distance classification test. Each time a reference speaker is chosen as having the minimum distance, its classification count is incremented by one. Acceptance and rejection thresholds can be defined in terms of statistical probabilities, where the strength of the decision can be specified according to the desired probability of false rejection, P[FR], and the probability of false acceptance, P[FA].

The sequential analysis decision procedure was applied to the results of a speaker identification experiment with no vowel recognition. The test vowel samples were processed in discourse order. The sequential decision results for test speaker HAN are illustrated graphically in Figure 3. The abcissa represents the number of samples tested, and the ordinate represents the classification count (number of minimum distance classifications). The two parallel lines represent the acceptance and rejection decision thresholds, which were computed according to the following specifications: P[FA]=.Ø5, P[FR]=.Ø5. The incremental lines represent the various reference speakers. When a reference accumulates a sufficient number of test samples, the acceptance threshold is crossed and an identification decision can be made at that time. In the case illustrated in Figure 3, the HAN reference reached the acceptance threshold after only 14 vowel samples, or after a time span of only 12 seconds.

As a result of applying sequential analysis, 17 out of 2Ø speakers tested were correctly identified at the specified error probability levels and one at a slightly larger error probability level. The amount of time to reach a decision ranged between 5.5 seconds and 9Ø seconds, with an average decision time of 29.4 seconds. For one of the test speakers the decision narrowed down to two possible references, with the true speaker being a member of the pair. Only one test speaker in twenty was falsely identified.

References

1) Pfeifer, Larry L. (1977). "An Interactive Laboratory System for Research in Speech and Signal Processing", submitted to IEEE Transactions on Acoustics, Speech, and Signal Processing.
2) Wald, Abraham (1952). Sequential Analysis, John Wiley and Sons, Inc., New York.

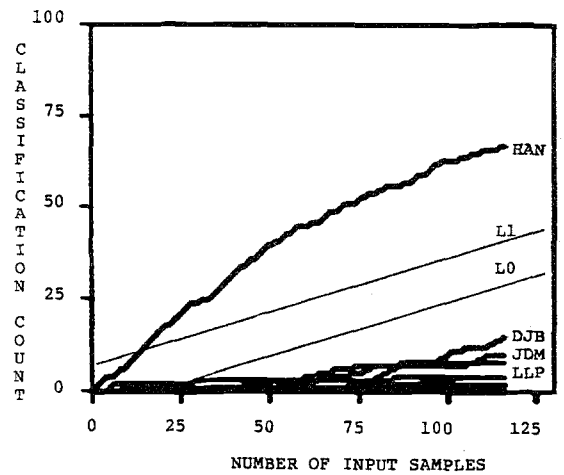

Figure 3. Plot of sequential analysis result for text-independent speaker identification without vowel recognition. Test speaker is HAN. L1 = acceptance threshold and LØ = rejection threshold.