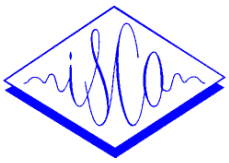

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEAKER SET IDENTIFICATION THROUGH SPEAKER GROUP MODELING

Jeff Kuo[†], *Chin-Hui Lee and Aaron E. Rosenberg*

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

ABSTRACT

Speaker identification has traditionally been performed by pattern matching an utterance from an unknown speaker to all the speaker models in the database. By clustering speakers into a fixed number of groups and performing speaker group identification instead, the computation required may be much less when the population of enrolled speakers becomes very large in size. This paper introduces the idea of speaker set identification and ways to efficiently accomplish speaker identification over a large population of speakers through speaker set identification and speaker clustering.

INTRODUCTION

Speaker identification is usually carried out by matching an unknown test utterance with individual speaker models. The speaker with the model that gives the best matching score is assigned as the identified speaker. However, for a large speaker population, this approach becomes impractical because the number of scores to be evaluated is prohibitively large. In this paper, we propose a novel approach which accomplishes *speaker set identification* [1] using speaker group models. First, based on a specified speech unit, the population of speakers is clustered into unit-dependent speaker groups. Since multiple units can be used for clustering speakers, we obtain multiple sets of speaker groupings each with a different partition of the speaker population. Each speaker in the population is therefore associated with a set of unit-dependent group memberships which we specify by *speaker group code*. For a given subset of units each speaker is associated with the intersection of the speaker groupings specified by the elements of the speaker group code. This intersection is referred to as the *speaker set* which generally contains more than one speaker. Speaker set identification is performed by matching an unknown test utterance with group models corresponding to the units representing the utterance. The speaker set that gives the best overall group matching scores is assigned as the identified speaker set. If the units used in clustering are acoustically distinct and the number of units is large enough, each speaker can often be uniquely identified using the corresponding speaker set. Compared with the number of individual speakers in a large population, there is usually only a small number of combined units and speaker groups. Therefore, speaker identification may be achieved efficiently using the proposed technique.

In the current study, we used an isolated digit database of 100 speakers consisting of 50 females and 50 males. The database was recorded over dialed-up telephone lines. 20 utterances for each digit were collected in five sessions for each speaker. Data from the first two sessions (8 tokens) were used for speaker clustering and speaker group model training. Data from the last three sessions (12 tokens each) were used for testing. This database has previously been used in other speaker recognition experiments and are described in [2, 3, 4]. In our experiments, each digit (10 of them in total) is considered as a distinct unit and for each digit, digit-dependent speaker clustering is performed to partition the set of 100 speakers into *eight* digit-dependent speaker groups. In our experiment, each speaker group is modeled by a continuous density hidden Markov model (HMM) with mixture Gaussian state observation densities. This results in a total of 80 group models compared to a total of 1000 digit models if no speaker grouping is performed. The computation saving here is only 80 versus 1000. It becomes more substantial if a large number of speakers are compared.

[†]Jeff Kuo is now with the Massachusetts Institute of Technology, Cambridge, MA 02139.

SPEAKER SETS AND SPEAKER GROUP CODES

For each speech unit w, the speaker population $S = \{X_1, X_2, \ldots, X_n\}$ can be clustered, based on the training utterances, into a set of g non-overlapping classes or groups $\{C_{w1}, C_{w2}, \ldots, C_{wg}\}$, where the group k of unit w is denoted as C_{wk}. The speech unit can be defined acoustically or linguistically, i.e., units either based on distinct acoustic features or based on linguistic definitions such as a word. Group identification consists of identifying the group of the unknown speaker using an unknown test utterance. In our study, we use digits as the speech units for group clustering which results in a set of 10 digit-dependent speaker groupings comprising 80 group models ($g = 8$). Let $\{j_1, j_2, \ldots, j_{10}\}$ be the indices of clusters associated with speaker q for the 10 digits. This is referred to as speaker q's *speaker group code* which is established in the training process.

Based on a speaker's test utterances for the speech unit w, we know that the speaker is one of the members of the group C_{wk} if the group has been correctly identified. Suppose a speaker q is to be identified based on an utterance containing a sequence of d digits $\{w_1, w_2, \ldots, w_d\}$. The intersection of the digit-specific groups, $\bigcap_{i=1}^{d} C_{w_i j_i}$, is referred to as the *speaker set* that contains the speaker q for the given utterance. It is noted that the size of the speaker set is a function of both the number of speech units used in the test utterances and the amount of acoustic correlations across the various speech units. As the number of different units increases, the number of speakers sharing the same speaker group code may be reduced thereby reducing the ambiguity in identifying the speaker set corresponding to an unknown utterance. Prominent clusters like the male and female groups will retain their separation over different units, reducing the number of distinct speaker group codes. However, if two speakers consistently exhibit great similarity across

10.21437/ICSLP.1992-372

all the units, then they will always have the same speaker group code and cannot be distinguished from each other. Ideally we would like to have the group membership differ across a small number of the chosen speech units, so that each speaker will eventually have his or her own unique speaker group code.

As an example, Table 1 shows the speaker group codes for speakers 24, 27, 94, 66 and 52 in the database, using the clusters found for each individual digit. Except for two pairs of speakers, $\{48, 46\}$ and $\{80, 28\}$, each speaker has its own unique speaker group code. Assuming the test utterance contains all 10 different digits, a unique speaker group code allows the speaker to be distinguished from others by the groups to which it belongs.

SPEAKER SET IDENTIFICATION

In the following we contrast speaker group identification with speaker set identification. In speaker group identification we identify the speaker group associated with each spoken utterance. Based on the identities of the groups, say $\{C_{w_1 8}, C_{w_2 5}, \ldots, C_{w_{10} 6}\}$, the identified speaker(s) is given as that with the speaker group code $85 \ldots 6$. However, suppose a group identification error occurs in one or more of the digits. The resulting speaker group code may actually correspond to that of another speaker, in which case a speaker identification error occurs. On the other hand, it may correspond to a speaker set which is empty.

Since this scheme is vulnerable to a group identification error in any single digit, a lower bound on the error rate is the maximum of the group identification error rates of the different digits. In the worst case, if the group identification errors across the digits are negatively correlated, i.e. each speaker has a group identification error for one digit and no others, each group identification error can separately contribute to one speaker identification error. Thus an upper bound for the speaker identification error rate may be the sum of all the group identification error rates. Therefore this scheme does not seem satisfactory.

In speaker set identification the log likelihood scores computed for group identification are summed across the multiple digits. The idea is that a particularly low score for one digit, which would have resulted in a group identification error, may be offset by a particularly high score, relative to a competitor, for a different digit. This scheme requires more computation than group identification because summations over all possible speaker codes are needed to obtain the best scoring set. However, the amount of computation is still considerably less than comparing an unknown utterance to all the speaker models in a large population.

An assumption made in the experiments is that for a particular sequence of multiple digit utterances from a speaker, the digits are known, or can be recognized perfectly and segmented by some speech recognizer. For each of the 10 different digits uttered by a speaker, the log likelihood scores against all 8 group models are computed and stored in a matrix. These scores are then summed up in 100 (actually 98, the number of unique speaker group codes among all the speakers for 10 digits) different ways, corresponding to the speaker group code of each speaker, and each summed score is assigned to the proper speaker. More specifically, if the speaker group code of a speaker q is $j_1 j_2 \ldots j_{10}$, then the speaker's score for an utterance $O = O_{w_1}, \ldots, O_{w_{10}}$ consisting of the digits $w_1, w_2, \ldots, w_{10}$ is the sum of the scores of each digit utterance O_{w_i} against the group model of the speaker group $C_{w_i j_i}$. The speaker with the largest score is taken as the identified speaker (set). Although there are 8^{10} different codes for the 10 digits, the maximum number of summations is 100, the size of the population.

SPEAKER SET IDENTIFICATION RESULTS

In the following we presents the results of performing speaker identification through the speaker set identification scheme described in the last part of the previous section. In addition to single digit combination, identification is also performed with multiple repetitions of each digit, e.g. 2 repetitions for a total of 20 tokens, 3 repetitions for a total of 30 tokens, up to 12 repetitions for a total of 120 tokens. The speaker set identification error is plotted in Figure 1 as a function of the number of test digits.

The error rate when 10 different digits are used is high, 12.25%, compared to 0.33% for speaker identification with individual speaker models. Even with 120 test tokens, 5 speakers are misidentified. To understand why an error can occur even with so much data, we list the speaker group codes of these 5 speakers and their successful competitors in Table 1. The number of competing speaker sets ranges from one for speakers 27, 52 and 66 to 7 for speaker 24.

Although these errors may not be indicative of errors in general, they give an idea of what some of the reasons for speaker set identification errors are. From speaker identification experiments, it has been known that speakers 27 and 24 are problem speakers with serious training and test token mismatch. Speaker 27 is also often considered an outlier during clustering. Thus a combination of poor clustering and modeling resulted in poor true group scores and contributed to the speaker identification errors. For the other three speakers, 94, 66, and 52, it is clear that the top competitors have speaker group codes which are too similar. For example, the only difference between the codes for speaker 52 and its competitor 72 is in the first digit. An error in group identification for the first digit in favor of group 8 instead of group 1 causes the speaker identification error. Perhaps such speakers as 52 can be reassigned to a different group for certain digits so that their codes are more different from potential competitors.

When a set of 10 different digits are used in speaker identification using individual speaker models, an error rate of 0.33% is possible. It would seem that speaker identification through speaker set identification has a rather poor identification performance compared to using individual speaker models. However, the amount of computation required is substantially less. The number of likelihood score computation depends on the number of group models regardless of the size of the speaker population. It is likely that the number of parameters used to specify the group models for each digit will remain fairly constant even when the number of speakers becomes very large. In addition, the identification performance may be improved incrementally with a tradeoff in time by using individual speaker models of the top n competitors in a second phase of identification, as discussed above. In this experimental study, using the top ten speakers, one can obtain an error rate of about 0.92%, which is close to 0.33% when all speaker models are included in pattern comparison.

DIFFERENT DIGIT COMBINATIONS

In the previous section, we considered speaker set identification when test trials contained all 10 different digits. In this section we explore the dependence of the speaker set identification error rate on the number of digits used. The group identification performance dramatically improves with the addition of different digits. One

might have expected then that speaker set identification should become more accurate when different digits are used.

Figure 2 shows the speaker set identification results as a function of the number of digits used. To give an idea of how different the relation can be, the results show two cases where the digits are arranged in two different orders. The one that starts with "four" and ends with "seven" is arranged in order of decreasing group identification error rates. Previous group identification experiments indicate that "four" is the digit with the biggest group identification error rate, and "seven" is the digit with the lowest. The one that starts with "seven" and ends with "four" is exactly the reverse, arranged in order of increasing error rates.

The top figure of Figure 2 shows the speaker set identification error rate as a function of the number of digits. In contrast to group identification, the error rate does not decrease monotonically. Rather, it increases rapidly at the beginning and then drops off gradually as the number of digits increases. The bottom figure shows the number of nonempty speaker sets as a function of the number of digits used. In either sequence of digits, there are eight sets when one digit is used and 98 sets when all ten digits are used.

In the top figure, when less than five digits are used, the error rates with the sequence of digits arranged in order of increasing group identification error rates are much higher than those with the sequence arranged in order of decreasing error rates. The digits "four," "six," "eight," and "two" do not seem to discriminate speakers as well as the digits "seven," "zero," "nine," and "three." With the introduction of the digit "nine" as the eighth digit in the sequence "$4, 6, 8, \ldots, 9, 0, 7$," the speaker set identification error rate dropped from 23.58% to 18.25%. If the digit "seven," which nominally has the lowest group identification error rate, were used here instead of "nine", the error rate would have been 20.58% instead of 18.25% for 8 digits. Thus it appears that group identification error rates by themselves cannot totally predict which combinations of digits will do best in set identification. The digit "nine" seems to have provided better complementary information than the digit "seven" to the other seven digits.

Next we address the question of why the speaker set identification error rate increases dramatically at first when multiple digits are used. The error rate when two digits are used is almost as large as the sum of the group identification error rates for the individual digits. For example, the sum of the group identification error rates for the digits "four" and "six" is 36.17% and for "seven" and "zero" is 18.92%, compared with the speaker set identification error rates of 31.42% and 16.83%, respectively. Because most of the speaker sets differ in their codes in only one place, a single group identification error easily leads to a speaker set identification error.

It is thus desirable to have speaker group codes which differ from each other by as many places as possible, i.e. have the minimum Hamming distance between any two codes as big as possible. The problem of finding such codes deserves some future study. Some digits may be suitable for quickly splitting the speakers into singletons initially, while other digits may be used to improve the identification error rate. In this way, a crude identification can be made with fewer digits, and the degree of confidence increased with every additional digit.

SUMMARY

This paper describes a simple paradigm for performing speaker identification using speaker group models. Although the speaker set identification error rate of 12.25% when 10 different digit tokens are used is poor compared with the baseline performance of 0.33%, there are two important advantages. First, the number of models per digit is reduced from 100 speaker models to 8 group models. Thus the amount of likelihood score computation needed is reduced. The required storage is also substantially reduced. Second, multiple candidate speakers can be selected to reduce the size of the speaker population to be compared and then individual candidate models can be used to perform 'true' speaker identification. By doing so, a performance of less than 1% error can be achieved.

The relatively high error rate in speaker set identification seems to be caused by the intrinsic problems in speaker clustering. In a separate study reported elsewhere [5], we found that the group identification error rate is rather high when speaker clustering beyond the two natural female and male groups is attempted. We do believe that speaker set identification can be enhanced if improved speaker clustering techniques can be incorporated in generating the speaker group codes and speaker sets.

REFERENCES

[1] A. E. Rosenberg and K. L. Shipley. Preliminary experiments on speaker recognition and classification coincident with speaker independent digit recognition. *Bell Laboratories Technical Memorandum*, June 1980.

[2] F. K. Soong, A. E. Rosenberg, B. H. Juang, and L. R. Rabiner. A vector quantization approach to speaker recognition. *AT&T Technical Journal*, 66(2):14–26, March–April 1987.

[3] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. ASSP*, 36(6):871–879, June 1988.

[4] A. E. Rosenberg, C. H. Lee, F. K. Soong, and M. McGee. Experiments in automatic talker verification using sub-word unit hidden Markov models. In *Proc. ICSLP-90*, Kobe, Japan, November 1990.

[5] J. Kuo. Speaker Clustering with Hidden Markov Models. *M.S. Thesis*, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, May 1992.

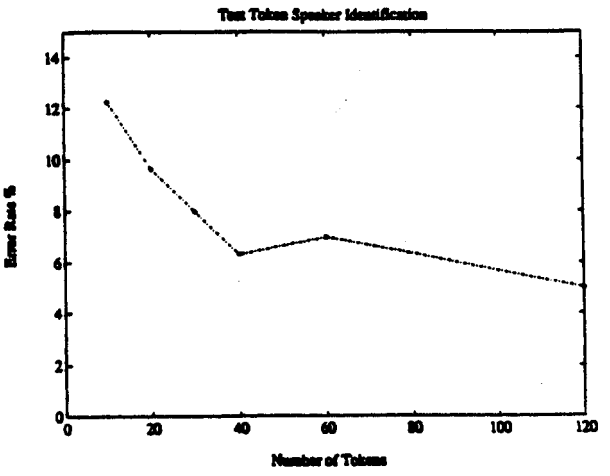

Figure 1: Speaker set identification results through combining group model scores of up to 12 repetitions of 10 different digits

Sp	Code	Sp	Code	Sp	Code	Sp	Code	Sp	Code
24	3812453451	27	2457241214	94	7638454477	66	3164163361	52	1612154447
48,46	3612424451	47	2456546218	58	7638424447	84	3164463351	72	8612154447
30	3612123456			56	8638454447				
80,28	3112163356								
62	3164423351								
22	3612163356								
32	8612454441								
8	8112134451								

Table 1: The speaker group codes of the five speakers misidentified, and the codes of the competing speakers who have higher scores.

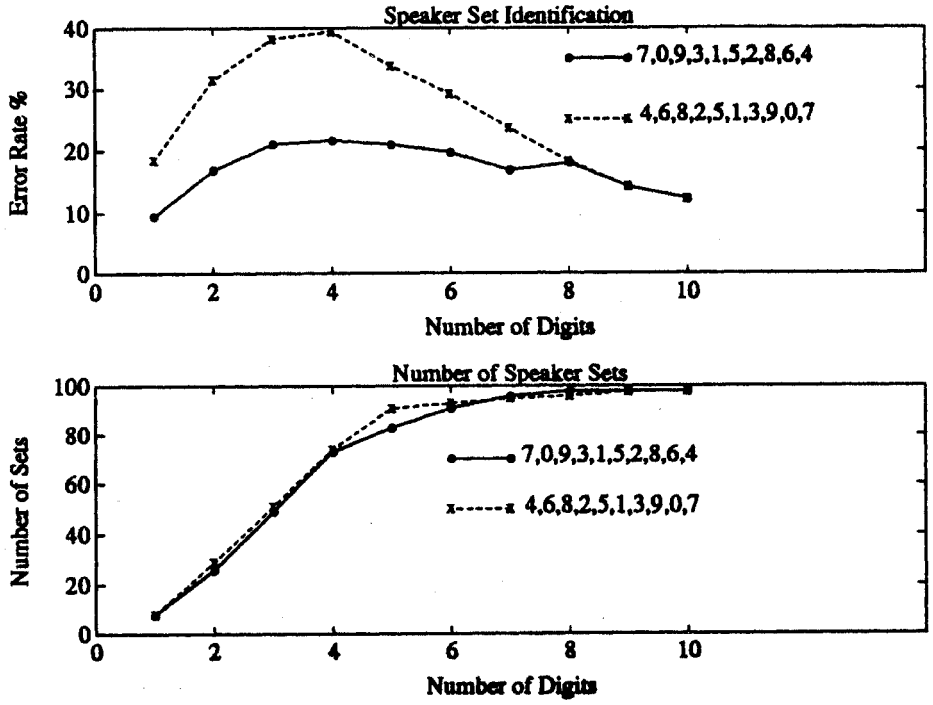

Figure 2: Speaker set identification error rate and the number of speaker sets as a function of the number of different digits used. The digits are arranged in two different orders, as shown in the legend in the bottom figure. For example, one of them is "4, 6, . . . , 7" so that a value of 5 on the x-axis refers to using the combination of the 5 digits "4," "6," "8," "2," and "5."