# A NEW METHOD OF TEXT-INDEPENDENT SPEAKER RECOGNITION

*Alan L. Higgins and Robert E. Wohlford*

*ITT Defense Communications Division*
*San Diego, CA 92131*

## ABSTRACT

Text-independent speaker recognition methods have been based on measurements of long-term statistics of individual speech frames. These methods are not capable of modeling speaker-dependent speech dynamics. In this paper, we describe a new method, based on template matching, that utilizes temporal information to advantage. The template-matching method performs text-dependent recognition as a special case. Performance of the template-matching method is compared with that of similar recently-developed methods.

## INTRODUCTION

Speaker recognition using short, prompted utterances is becoming practical because of the high accuracy that can be obtained. When the text of the utterance is unknown, however, accuracy is much lower. Further, existing algorithms for text-independent recognition are completely different than those used for text-dependent recognition. In this paper, we describe how a text-dependent algorithm can be extended to text-independent recognition with improved performance.

Text-dependent recognition systems are based on the principle of template matching. In a training session, each user donates templates for all the words in a limited vocabulary. Recognition is then accomplished by prompting the speaker to utter a phrase consisting of words from the vocabulary. A continuous speech recognition (CSR) system matches the phrase in turn with each user's template set. The recognition decision is a function of the resulting phrase-match scores.

In the algorithm described above, syntactic constraints can be applied within the CSR recognizer to force it to match the correct sequence of word templates. This causes knowledge of the specific text to be used, as well as knowledge of the vocabulary. Interestingly, doing so does not improve recognition performance. It appears. then. that the major difficulty in applying the template-matching method to text-independent recognition is not that the text of the utterance is unknown, but that the vocabulary is unlimited. Because of the unlimited vocabulary, it is impossible for users to donate templates for all the words that could be encountered in a conversation. However, if templates are derived from speech segments shorter than words, the problem of approximating an utterance of unknown text is simplified.

## THE NEW METHOD

The new method performs text-independent recognition using templates of the type described above. Template sets are required for each user or candidate speaker. A user's template set is derived by an automatic enrollment procedure using a sample of that person's conversational speech. A block diagram of the enrollment and recognition procedures is shown in Figure 1. These procedures are described in the following two sections.

### Enrollment

The purpose of enrollment is to produce a set of templates that characterizes the acoustical space used by an individual speaker. We call these templates "filler templates" because they can be used to fill in or match speech for which word templates are not available[1].

Filler templates are short acoustic segments that need not have any linguistic identity such as phonemes, diphones, or syllables. They represent the speaker in the sense that they can be concatenated to approximate acoustically any new utterance by the same speaker. To simplify the enrollment procedure and to give every speaker equal treatment, our experiments to date have used template sets in which all speakers have an equal number (N) of templates and all templates have an equal number (M) of frames.

Filler templates were derived by applying the "K-means" cluster averaging procedure to the enrollment utterance. The clustering criterion to be minimized was the mean squared difference, after time warping, between the enrollment utterance and the best template sequence. After three iterations of the algorithm, this criterion was reduced to about half its starting value.

A covering algorithm was used to generate the "starting points" for clustering. The covering algorithm scanned the enrollment speech, adding new templates to the template set until every M-frame segment of enrollment speech was within a specified distance of the nearest template. The specified distance was chosen to give more than N templates. Exactly N templates were


CH2243-4/86/0000-0869 $1.00 

then selected as follows. A matrix encoder encoded the enrollment speech using the full template set as the codebook. A histogram was maintained of the frequency of usage of the templates. The N most frequently-used templates were selected.

Recognition

The speaker of an utterance is recognized by matching the utterance with user's template sets using a CSR system. No syntactic constraints are imposed in the matching. The recognition algorithm can be used with either text-dependent or text-independent speech material, changing only the templates. Whereas whole-word templates can be used if the text is controlled, filler templates are used in the text-independent case.

To identify the speaker from a closed set of candidates, the decision procedure in Figure 1 is to choose the speaker with the lowest phrase-match score. To verify a claimed identity, a simple decision procedure is to compare the phrase-match score for the claimed speaker with a threshold, verifying the claim if the score is the lower. The main difficulty with this method is that match scores are influenced by both the speaker and the phonetic content of the verification utterance. To address this problem, we developed a scoring method called "likelihood scoring", which is used as the decision procedure in Figure 1 for speaker verification. This method forms a likelihood score that is a function of the phrase-match score for the claimed speaker and the phrase-match scores for a set of other speakers. The likelihood score is an approximation to the likelihood ratio, or the ratio of the likelihood of the observed utterance assuming the claimed speaker to the likelihood of the observed utterance assuming a different speaker. The likelihood score is compared with a threshold to make the verification decision. Figure 2 is a block diagram of the likelihood scoring method. The template donors other than the claimed speaker make up what is called the "imposter set". In our simulations, every verification trial used an imposter set formed from the whole user set by excluding the true speaker and the claimed speaker. This procedure was used because of the small number of speakers used in the test database. The value of the constant alpha in Figure 2 lies between 0 and 1 and is chosen experimentally. A value of 0.5 was used in the testing described below. Likelihood scoring can be used in both text-dependent and text-independent verification.

## SIMILAR RECENTLY-DEVELOPED METHODS

The new template-matching method is similar to a method based on vector quantization described at ICASSP 85 by Soong, et al.[2]. The vector-quantization method is equivalent to the template-matching method when the length of the templates is 1 frame.

The template-matching method is also similar to a method based on matrix quantization described at ICASSP 85 by Burton[3]. In that study, the matrix-quantization method was applied to the problem of isolated word recognition. It could be used for speaker recognition by taking the average encoding error as a measure of speaker distance. CSR template matching and matrix quantization are similar in that they both perform a segmentation and labeling of the input speech in terms of a set of multiple-frame patterns. The main difference between the two is that matrix quantization does not allow time warping. The template-matching method extends both the vector-quantization and matrix-quantization methods.

## DATABASE

We tested the template-matching method on a standard government database of conversational speech from 11 male speakers[4]. For each speaker, Session "6" was used for enrollment and session "7", recorded one week later, was used for testing. After removing long pauses, the duration of each conversation was about 100 seconds. The recordings were made in a quiet room using a high-quality microphone, and digitized at 8000 samples per second. Ten cepstral coefficients per 10-ms frame were derived from LPC-12 analysis using the autocorrelation method.

The test conversations were divided into 2.5-second segments. In our experiments, recognition accuracy was measured for individual and multiple 2.5-second segments. To measure recognition accuracy for, say, 10-second utterances, the phrase-match scores within groups of 4 consecutive segments were averaged.

## EXPERIMENTS AND RESULTS

We measured the performance of the template-matching method under a variety of conditions, and compared the template-matching method with the recently-developed methods discussed above. As a baseline for the experiments, we took the following system. Speaker-dependent covering analysis was used, generating from 200 to 300 covers per speaker. Fifty filler templates per speaker were derived by cluster averaging. Each template contained eight 10-ms frames. Unless otherwise indicated, the results reported in this section are for closed-set speaker identification.

Figure 3 compares the recognition error rate of the baseline system described above with that of a previously-developed speaker-recognition system. This previous system was based on measurement of distances between the long-term means of parameters of the test and enrollment utterances using a weighted Euclidean distance[5]. The error rates, in percent, are shown as functions of the length of the test utterance. Curve A represents the previous system, curve B represents the template-matching system, and curve C represents a combination of the A and B systems in which individual speaker scores for the two systems were linearly combined. The template-matching system is clearly superior to the previous system for all test lengths. The com-

bined system performs better than either individual system for short utterances.

Although we did not attempt to optimize the template length, we did perform an experiment using single-frame templates. To do this, the CSR phrase matcher was replaced with a vector quantizer, resulting in a system similar to that of the vector-quantization discussed above.[2]. The performance of this system is shown in Figure 4. For 2.5-second test utterances, single-frame templates perform as well as 8-frame templates. As utterance length increases, the accuracies diverge. The accuracy using single-frame characters reaches a plateau at about 4 percent, while the accuracy using 8-frame characters continues to improve.

In another experiment, we replaced the CSR phrase matcher of the baseline system with a matrix quantizer in both the enrollment and recognition algorithms. Two types of matrix quantization were compared, following the work of Burton[3]. The first, called "jumping", encodes contiguous nonoverlapping 8-frame segments of the input speech. The second, called "sliding", encodes 8-frame segments that are advanced only one frame at a time. Both matrix-quantization methods give performance inferior to that of the baseline system as shown in Figure 5. As in the Burton study, the sliding method is better than the jumping method. The computational complexity of the sliding method is roughly equal to that of the baseline system.

Finally, we measured the performance of the system for speaker verification. In the test, each of the 11 speakers served in turn as "valid users" and as imposters for the other 10 speakers. Figure 6 shows equal-error rates for speaker verification based on phrase-match scores (Curve A) and likelihood scores (Curve B).

## DISCUSSION

The template-matching method is dramatically superior to the previous method tested, particularly for longer test utterances. We believe the reason is that template matching compares segments of the test utterance with acoustically-similar segments of the enrollment utterances. In doing so, it takes a step toward bridging the gap between text-independent and text-dependent speaker recognition.

Cluster averaging is an essential part of the enrollment algorithm of the template-matching method. Performance of the method is relatively insensitive to the choice of starting points for clustering. We found that performance similar to that of the baseline system was obtained when starting points for all speakers were derived from a single speaker-pooled covering analysis, instead of separate speaker-dependent analyses.

The improved performance of 8-frame templates over single-frame templates is evidence that the dynamics of speech spectra within 80-ms segments is an important source of speaker information. The 8-frame templates, however, benefit from greater lengths of both training and test data.

In matching multiple-frame templates, it appears to be desirable to allow for differences in rate of speech by using dynamic time warping.

Speaker verification may be the most important application of the template-matching method. The likelihood scoring method substantially improves verification performance. The ability of a single algorithm to perform both text-dependent and text-independent verification is important in certain applications such as secure telephones. In such applications a user continues to speak after initially gaining access through text-dependent verification. Text-independent verification can then monitor the speech to ensure that a different speaker does not take over. The method described in this paper is particularly suited to such applications because of the economy afforded by combining the two functions.

## SUMMARY

We have described a new method of text-independent speaker recognition that is an extension of a method used for text-dependent recognition. The new template-matching method compares segments of the test utterance with acoustically-similar segments of the enrollment utterances. It characterizes speech dynamics of individual speakers. For 10-second test utterances, the identification error rate of the template-matching method is an order of magnitude lower than that of a previously-developed method.

The template-matching method has been applied to the speaker verification task using a new scoring method called likelihood scoring. Both text-dependent and text-independent speaker verification are performed using a single algorithm.

## References

1. A. Higgins and R.E. Wohlford, "Keyword Recognition Using Template Concatenation," *Proc. Internatl. Conf. Acoust., Speech, and Sig. Proc.*, Tampa, FL, March 1985.
2. F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A Vector Quantization Approach to Speaker Recognition," *Proc. ICASSP 85*, vol. 1, pp. 387-390, Tampa, FL, 1985.
3. D. K. Burton, "Applying Matrix Quantization to Isolated Word Recognition," *Proc. ICASSP*, vol. 1, pp. 29-32, Tampa, FL, 1985.
4. J. D. Markel, B. T. Oshika, and A. H. Gray Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 330-337, 1977.
5. E. H. Wrench, "A realtime implementation of a text independent speaker recognition system," *Proc. IEEE Internat. Conf. Acoust., Speech and Signal Process.*, no. 1, pp. 193-196, 1981.

then selected as follows. A matrix encoder encoded the enrollment speech using the full template set as the codebook. A histogram was maintained of the frequency of usage of the templates. The N most frequently-used templates were selected.

Recognition

The speaker of an utterance is recognized by matching the utterance with user's template sets using a CSR system. No syntactic constraints are imposed in the matching. The recognition algorithm can be used with either text-dependent or text-independent speech material, changing only the templates. Whereas whole-word templates can be used if the text is controlled, filler templates are used in the text-independent case.

To identify the speaker from a closed set of candidates, the decision procedure in Figure 1 is to choose the speaker with the lowest phrase-match score. To verify a claimed identity, a simple decision procedure is to compare the phrase-match score for the claimed speaker with a threshold, verifying the claim if the score is the lower. The main difficulty with this method is that match scores are influenced by both the speaker and the phonetic content of the verification utterance. To address this problem, we developed a scoring method called "likelihood scoring", which is used as the decision procedure in Figure 1 for speaker verification. This method forms a likelihood score that is a function of the phrase-match score for the claimed speaker and the phrase-match scores for a set of other speakers. The likelihood score is an approximation to the likelihood ratio, or the ratio of the likelihood of the observed utterance assuming the claimed speaker to the likelihood of the observed utterance assuming a different speaker. The likelihood score is compared with a threshold to make the verification decision. Figure 2 is a block diagram of the likelihood scoring method. The template donors other than the claimed speaker make up what is called the "imposter set". In our simulations, every verification trial used an imposter set formed from the whole user set by excluding the true speaker and the claimed speaker. This procedure was used because of the small number of speakers used in the test database. The value of the constant alpha in Figure 2 lies between 0 and 1 and is chosen experimentally. A value of 0.5 was used in the testing described below. Likelihood scoring can be used in both text-dependent and text-independent verification.

## SIMILAR RECENTLY-DEVELOPED METHODS

The new template-matching method is similar to a method based on vector quantization described at ICASSP 85 by Soong, et al.[2]. The vector-quantization method is equivalent to the template-matching method when the length of the templates is 1 frame.

The template-matching method is also similar to a method based on matrix quantization described at ICASSP 85 by Burton[3]. In that study, the matrix-quantization method was applied to the problem of isolated word recognition. It could be used for speaker recognition by taking the average encoding error as a measure of speaker distance. CSR template matching and matrix quantization are similar in that they both perform a segmentation and labeling of the input speech in terms of a set of multiple-frame patterns. The main difference between the two is that matrix quantization does not allow time warping. The template-matching method extends both the vector-quantization and matrix-quantization methods.

## DATABASE

We tested the template-matching method on a standard government database of conversational speech from 11 male speakers[4]. For each speaker, Session "6" was used for enrollment and session "7", recorded one week later, was used for testing. After removing long pauses, the duration of each conversation was about 100 seconds. The recordings were made in a quiet room using a high-quality microphone, and digitized at 8000 samples per second. Ten cepstral coefficients per 10-ms frame were derived from LPC-12 analysis using the autocorrelation method.

The test conversations were divided into 2.5-second segments. In our experiments, recognition accuracy was measured for individual and multiple 2.5-second segments. To measure recognition accuracy for, say, 10-second utterances, the phrase-match scores within groups of 4 consecutive segments were averaged.

## EXPERIMENTS AND RESULTS

We measured the performance of the template-matching method under a variety of conditions, and compared the template-matching method with the recently-developed methods discussed above. As a baseline for the experiments, we took the following system. Speaker-dependent covering analysis was used, generating from 200 to 300 covers per speaker. Fifty filler templates per speaker were derived by cluster averaging. Each template contained eight 10-ms frames. Unless otherwise indicated, the results reported in this section are for closed-set speaker identification.

Figure 3 compares the recognition error rate of the baseline system described above with that of a previously-developed speaker-recognition system. This previous system was based on measurement of distances between the long-term means of parameters of the test and enrollment utterances using a weighted Euclidean distance[5]. The error rates, in percent, are shown as functions of the length of the test utterance. Curve A represents the previous system, curve B represents the template-matching system, and curve C represents a combination of the A and B systems in which individual speaker scores for the two systems were linearly combined. The template-matching system is clearly superior to the previous system for all test lengths. The com-