Closed-Set Speaker Identification Based on a Single Word Utterance: An Evaluation of Alternative Approaches

G.R.Dhinesh, G.R.Jagadeesh, T.Srikanthan
Nanyang Technological University, Centre for High Performance Embedded Systems
50 Nanyang Drive, Singapore 637553
grdhinesh@ntu.edu.sg; asgeorge@ntu.edu.sg; astsrikan@ntu.edu.sg

Abstract **-**The problem of closed-set speaker identification based on a single spoken word from a limited vocabulary is relevant to several current and futuristic interactive multimedia applications. In this paper, we evaluate the effectiveness of several potential solutions using an isolated word speech corpus. In addition to evaluating the text-dependent and text-constrained variants of the Gaussian Mixture Model (GMM) based speaker identification system, we also propose and evaluate a method that seeks to improve the speaker identification performance by efficiently segmenting words into sub-word units and grouping similar sub-words together. For the task of identifying a speaker among 16 speakers uttering any word from a 20-word vocabulary, the text-dependent speaker identification method achieved a speaker identification rate of over 99%. However, this is conditional upon the test word being correctly identified by a speech recognition module. While the text-constrained speaker recognition system produced a speaker identification rate of over 96% for the 20-word vocabulary, better results were obtained when the vocabulary size was reduced. The best result achieved by the proposed method based on sub-word grouping has been found to be almost identical to that of the text-constrained speaker identification method.

Keywords**:** Speaker recognition, closed-set identification, speech database, voice-based interaction

1. Introduction

There are a number of emerging interactive applications that could potentially benefit from the ability to identify a user, among a closed set of users, based on a single spoken word from among a limited set of words. Examples of such applications include voice-based multiplayer gaming and intelligent devices that respond to commands in a user-dependent manner. The short, typically sub-second, duration of the test utterance rules out the use of conventional text-independent speaker recognition systems, where test utterances between 10 and 30 seconds are considered reasonable [1]. This paper aims to evaluate several potential alternatives to text-independent speaker recognition for the above problem including a novel method proposed by the authors.

Since we are dealing with a limited vocabulary of words, it is not infeasible to employ the text-dependent speaker recognition approach, where the same text is used for training (enrolment) and testing (identification). It is well-established that Speaker recognition using text-dependent speech achieves much better performance than using text-independent speech because in the former, comparisons are only made between how speakers produce certain specific sounds [2]. However, the text-dependent approach has a few disadvantages. For instance, in the applications targeted by our research, a user can potentially utter any word from the given set of words and therefore text-dependent speaker models need to be prepared for each word in the vocabulary. In such situations, a speech recognition system is required to identify the spoken word so that the corresponding speaker model can be used for speaker identification. This makes the speaker identification process dependent on the correctness of speech recognition.

Text-constrained speaker recognition overcomes the above disadvantages by providing a limited degree of text-independence. In a text-constrained system, a single speaker model is generated for each speaker using all the words present in a limited vocabulary [2]. During the identification phase, the user utters one of the words from the given vocabulary, which is scored against the speaker model. In this paper, apart from evaluating the text-dependent and text-constrained speaker recognition methods on an

isolated word speech database, we present and evaluate a potential improvement to the latter. The proposed method efficiently segments words into sub-word units, groups similar sub-words together and models each group separately so that the best-matching model can be used for each sub-word unit in the test utterance.

In all the methods described in this paper, the speaker models have been generated using Gaussian Mixture Model (GMM), which has been established as the most successful technique for speaker recognition [3]. In line with the standard practice, Mel Frequency Cepstral Coefficients (MFCC) are used as features. 20 dimensional MFCC feature vectors are calculated from a 20 ms speech window progressing at a rate of 10 ms. All the experiments have been carried out using the TI20 section of the TI46 database [4], which contains several utterances of 20 isolated words from 16 speakers (8 male and 8 female). The words in the database are ZERO, ONE, TWO, THREE, FOUR, FIVE, SIX, SEVEN, EIGHT, NINE, ENTER, ERASE, GO, HELP, NO, RUBOUT, REPEAT, STOP, START and YES. The database is organized into two sessions. One session with a total of 5120 words is used for training and test utterances are taken from another session with 3200 words.

In the following two sections, we evaluate the text-dependent and text-constrained speaker identification methods on the TI46 speech corpus and analyse the results. In Section 4, we present and evaluate the proposed sub-word based improvement to text-constrained speaker identification. In Section 5, we discuss the findings and conclude the paper.

2. Text-Dependent Speaker Identification

Text-dependent speaker recognition is typically employed in applications such as access control, where a fixed phrase is used as a password and the registered speaker models are trained using only the utterances of that phrase. However, in applications where the test utterance can be one of many permitted words, multiple speaker models, trained using each word in the vocabulary, should be created for each speaker. In our experimental setup based on the TI46 database with 16 speakers and 20 words, we have trained a total of $16 \times 20 = 320$ speaker models. Each test utterance is scored against all the 16 speaker models corresponding to the word according to the procedure given in [3] in order to obtain a measure of likelihood. Since we are concerned with the closed set identification problem, we do not normalize the likelihood against a background model. Instead, the speaker whose model yields the highest likelihood is identified as the source of the test utterance.

Table 1: Text-dependent speaker identification

No. of training utterances	Speaker Identification rate (%)			
	No. of GMM components: 4	No. of GMM components: 8	No. of GMM components: 16	No. of GMM components: 32
5	98.02	97.93	97.08	96.73
6	98.43	98.68	97.96	97.80
7	98.52	99.00	98.61	98.61
8	98.68	99.02	98.96	98.80
9	98.74	99.09	99.05	98.90
10	98.93	99.28	99.24	99.02

In the experiments, we varied the number of components in the GMM and number of instances of a word used to train a speaker model. Results presented in Table 1 indicate that the best speaker identification rate is achieved when 8 GMM components are used. Quite predictably, the identification rate improves marginally with the increase in the number of word utterances used for training. While the text-dependent approach achieves a high speaker identification rate of over 99%, it needs to be noted that in the experiments, the identity of the words used for testing is correctly known. However, in practice, a speech recognizer would be used to first identify the word before performing text-dependent speaker identification as illustrated in Figure 1. In such a situation, speech recognition errors are likely to degrade the speaker identification rate.

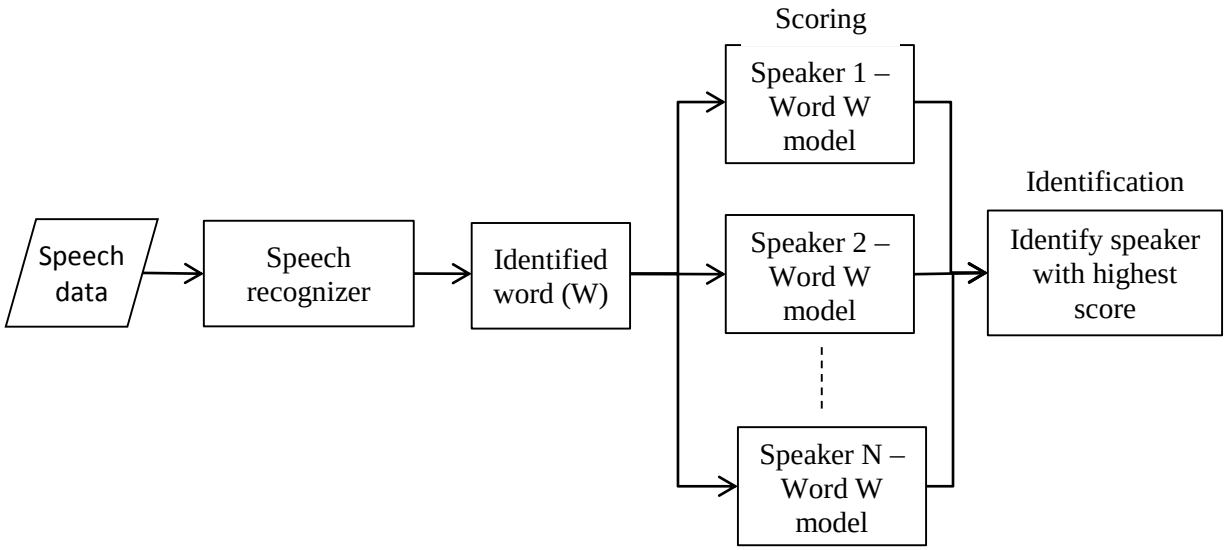

Figure 1: Text-dependent speaker identification for a limited vocabulary

3. Text-Constrained Speaker Identification

We have implemented a text-constrained speaker identification system by training a single model for each speaker by considering all the instances of all the words in the vocabulary spoken by him or her. During testing, the input word is scored against all the 16 speaker models to identify the one with the highest likelihood. In order to study the effect of the vocabulary size on the speaker identification performance, we have experimented with the following 3 different sets of words.

Table 2: Text-constrained speaker identification with different vocabulary sizes

Vocabulary	Vocabulary size	Speaker identification rate (%)			
		No. of GMM components: 8	No. of GMM components: 16	No. of GMM components: 24	No. of GMM components: 32
Set 1	20	88.54	93.03	94.98	96.60
Set 2	10	90.71	93.66	97.05	97.55
Set 3	5	95.98	97.86	98.49	98.74

a) Set 1 includes all the 20 words in the TI46 database
b) Set 2 includes only the 10 digits.
c) Set 3 is limited to the following 5 words: ENTER, ERASE, GO, HELP and NO.

Having more words in the vocabulary means that a higher amount of speech data is available for training the speaker model. For instance, when Set 1 is used as the vocabulary, each speaker is trained with 200 (20 words × 10 instances) words. Despite using more speech data for training, an increase in the vocabulary size has been found to cause a marginal decrease in the speaker identification rate as shown in Table 2. The best results were obtained when 32 components were used in the GMM. For Set 1, which includes all the 20 words in the database, the achieved speaker identification rate of 96.6% is lower than the identification rate obtained by the text-dependent approach.

4. Speaker Identification with Sub-Word Grouping

We seek to improve the robustness of the test-constrained speaker recognition system by segmenting and grouping the speech data into different sound classes and confining the speaker recognition process to comparisons within specific sound classes. This approach has been explored by other researchers [5][6][7], who used phonemic decoders to segment and group speech data into broad phonetic classes such as vowels, fricatives and plosives. Unlike these methods, we envisage a system that does not require a phoneme decoder, which typically relies on annotated transcriptions of a large amount of speech data. We are inclined towards a setup in which the utterances are segmented into sub-word units without using any linguistic knowledge. In an early work that presented a speaker recognition system based on Hidden Markov Models (HMM) of sub-word units, it was shown that there is only a small difference in performance between sub-word units formed with and without phonetic transcriptions of the utterances [8]. Although other researchers [9][10] have employed a language-independent segmentation method that does not require phonetic transcriptions, it is a data-driven method whose performance depends on the amount of training data and its effectiveness for speaker recognition has only been demonstrated on conversational speech databases. In contrast, we intend to investigate simple segmentation methods that do not require any training data or complex speech processing tools.

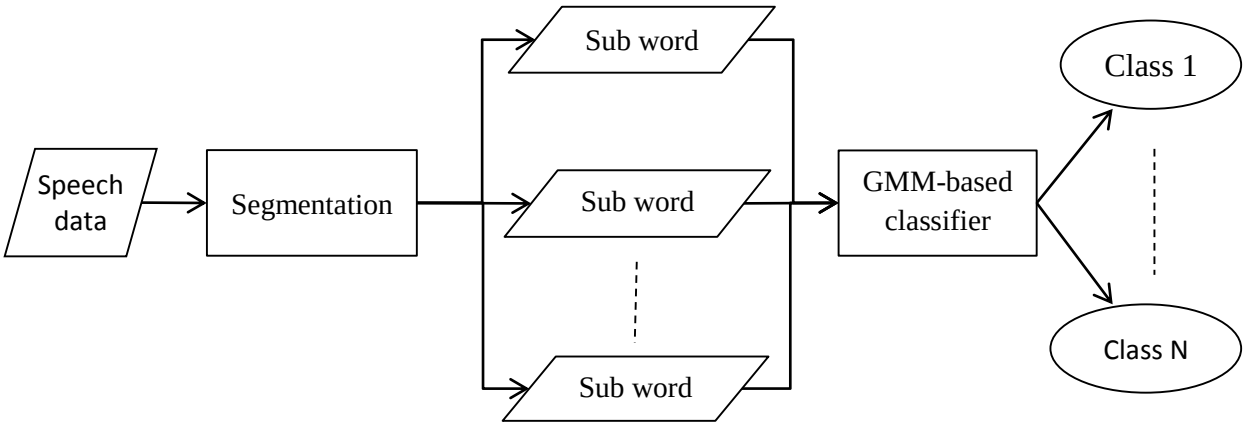

Figure 2: Overview of segmentation and sub-word grouping

An overview of the process of segmenting speech and grouping the sub-word units is shown in Figure 2. Each word in the speech data is segmented into a number of sub-word units using one of the two segmentation methods given below.

Equal segmentation: In this method, each word is segmented into a given number of equal segments. The number of sub-word units per word is predefined and is the same for all the words in the

system. If the number of feature vectors in a word cannot be divided equally among the sub-word units, then the length of the last unit is allowed to be slightly more than the rest of the units.

Fixed-length segmentation: In this method, all the sub-word units have a predefined, fixed length. Each word is divided into variable number of overlapping fixed-length segments. Hence the number of segments in a word varies according to the length of the word.

We make use of a GMM as a classifier to group the sub-word units into a number of sound classes. To classify the sub-word units into *C* sound classes, we use a classifier GMM with *C* mixture components. This GMM is intended to cover the global characteristics and hence it is trained by taking all the words in the system uttered by all the speakers. Each of the *C* components in the classifier GMM is intended to represent one sound class.

During speaker modelling, each of the sub-word units from a speaker is associated with one of the classes using the classifier GMM. This is done by scoring the feature vectors within the segment against each of the components in the classifier GMM and identifying the component that yields the highest score. Once all the sub-word units from the enrolment speech data is grouped into different classes, all the speech data within each class is trained separately to prepare a separate speaker model corresponding to that class.

During testing, each sub-word unit in the test word is associated with a sound class using the classifier GMM as described above. Subsequently, the each sub-word unit is scored only against the speaker models corresponding to that class. The overall matching score of a test word against a speaker's model is the sum of the scores obtained by all the sub-word units in it.

Table 3: Speaker identification with equal segmentation and sub-word grouping

No. of segments per word	Speaker identification rate (%)		
	No. of classes: 2	No. of classes: 4	No. of classes: 6
2	93.11	84.61	78.02
3	94.87	89.06	83.41
4	95.82	92.74	87.48
5	95.63	94.29	91.58
6	96.61	95.61	92.45

Table 4: Speaker identification with fixed-length segmentation and sub-word grouping

Length of the sub-word unit (No. of feature vectors)	Speaker identification rate (%)		
	No. of classes: 2	No. of classes: 4	No. of classes: 6
10	96.62	96.03	93.34
12	96.34	95.95	92.01
14	96.32	95.89	90.85
16	95.65	92.84	87.45
18	95.02	90.44	85.27

Two variants of the proposed speaker identification scheme, corresponding to the two segmentation methods, have been evaluated on the TI46 database. Tables 3 and 4 show the effect of the number and length of the sub-word units on the speaker identification rate. The results suggest that a higher identification rate is obtained when smaller sub-word units are used. It can perhaps be inferred that small segments are associated with distinct elementary sounds and hence lend well for being grouped into sound classes that are distinguishable from each other. However, the best identification rate achieved by the proposed method is almost identical to that of the text-constrained system.

5. Conclusions

The problem of performing closed-set speaker identification based on a single word utterance from a limited vocabulary has not received sufficient attention in the literature. In this paper, we have evaluated the feasibility of several alternative methods for the above problem using an isolated word speech database. While the text-dependent speaker identification system achieves the highest identification rate, such an approach sometimes requires an unwieldy number of speaker models to be prepared depending upon the number of permitted words. Its performance is also dependent on the correctness of the speech recognition front end used to identify the test words. Text-constrained speaker identification system, which only requires a single speaker model to be prepared using all the permitted words, offers some advantages over text-dependent systems. However, the performance of the former marginally deteriorates with an increase in the vocabulary size. We have proposed and evaluated a modification to the text-constrained speaker identification system based on the concept of efficiently segmenting words into sub-word units and grouping similar sub-words together. However, the results show that the proposed modification does not produce any noticeable improvement in the identification rate. Overall, all the methods evaluated in this paper achieve a speaker identification rate of over 96% for a closed set of 16 speakers and a vocabulary of 20 words.

References

[1] Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier S., Merlin, T., Ortega-Garcia, J., Petrovska, D., Reynolds, D. A. “A tutorial on text-independent speaker verification”, EURASIP Journal on Applied Signal Processing, pp. 430-451, 2004.

[2] Sturim, D.E., Reynolds D.A., Dunn R.B., Quatieri T.F. "Speaker Verification using Text-Constrained Gaussian Mixture Models", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 677-680, May 2002.

[3] Reynolds, D.A., Rose, R.C. “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83, 1995.

[4] Specification of TI46 Speech Database: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S9 (Accessed 22 May 2013).

[5] Koolwaaij, J., De Veth, J. "The use of broad phonetic class models in speaker recognition." Proceedings of the International Conference on Spoken Language Processing (ICSLP), 1998.

[6] Fatima, N., Zheng, T. F. “Syllable category based short utterance speaker recognition”, In Audio, Language and Image Processing (ICALIP), (pp. 436-441), July 2012.

[7] Zhang, C., Wu, X., Zheng, T. F., Wang, L., Yin, C. “A K-phoneme-class based multi-model method for short utterance speaker recognition”, In Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific (pp. 1-4), December 2012.

[8] Rosenberg, A. E., Lee, C. H., Soong, F.K. “Sub-Word Unit Talker Verification Using Hidden Markov Models,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 269-272, April 1990.

[9] Petrovska-Delacrétaz, D., Cernocky, J., Hennebert, J., Chollet, G. “Segmental Approaches for Automatic Speaker Verification”, Digital Signal Processing, Volume 10, Issues 1–3, pp. 198-212, January 2000.

[10] El Hannani, A., Petrovska-Delacretaz, D. “Improving speaker verification system using ALISP-based specific GMMs”, Audio-and Video-Based Biometric Person Authentication. Springer Berlin Heidelberg, pp. 580-587, 2005.