Latent Semantic Analysis for Speaker Recognition

A. Nayeeemulla Khan and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
Email: {nayeem,yegna}@cs.iitm.ernet.in

Abstract

There exists certain traits specific to a speaker that help in easy identification of the speaker among a familiar set of speakers. These include certain dis-fluencies, and mannerisms like stress for certain words, frequent usage of certain phrases, manner of pronunciation and back channels. The focus of this paper is identification of a speaker using such idiolectic traits in conversational speech. Every normal conversation by a speaker contains his idiolectic signature. A model is developed in the latent semantic analysis framework to capture this signature. The similarity of the idiolectic signature in the test utterance to that captured by the model is used to hypothesise the target speaker. The technique is demonstrated for the NIST 2003 extended data task.

1. Introduction

In conventional speaker recognition studies, short-time acoustic features are extracted from the speech signal and Gaussian mixture models or neural network models are trained to estimate the distribution of the feature vectors in higher dimensional space [1][2]. The advantage of such modeling is its simplicity and robustness. These techniques make no effort to capture the higher level knowledge present in speech. Use of prosodic information like pitch to identify speakers [3] and augmenting the GMM scores with prosodic and lexical information have been tried [4]. A similar task is that of authorship attribution, where the objective is to establish the authorship of anonymous or doubtful texts. Here processing of the text documents is carried out at the level of a word. The classic investigation is that of the *Federalist papers* written in 1787-1788 by Alexander Hamilton , John Jay and James Madison, to persuade the citizen of the New York state to ratify the US constitution. Twelve of these papers are of disputed authorship, said to have been written either by Madison or Hamilton. Statistical inference was used to conclude that the papers were written by Madison [5]. This has been extended by using a different set of function words [6]. Similar work has been done using style markers [7] and SVMs [8].

Humans are able to identify familiar speakers based on their idiolectic traits. Sometimes a distinct sound pattern from the speaker may be sufficient to identify the speaker. This is due to the presence of the speakers idiosyncrasies that the listener has learnt over time. With the availability of longer conversation data like switchboard corpus, it has been possible to capture the dis-fluencies and idiolectic characteristics of a speaker using an idiolectal language model [9]. Here a bigram language model was developed using transcriptions of each speaker's training data. Every test utterance was scored against these models.

This paper uses the concept of latent semantic analysis (LSA) to capture the idiolectic traits of speakers. In information retrieval, latent semantic indexing is a popular tool used to retrieve text documents based on their semantic similarities. The set of semantic concepts present in the documents are captured by means of the co-occurrences of words in a document using a matrix (W), with columns representing documents and rows representing words (terms). Singular value decomposition (SVD) is applied on W to compress and project the large dimensional vectors representing the term frequencies in documents onto a smaller continuous space of semantic concepts [10][11]. A test document/query is framed in the form of a term frequency vector like a column of W. The similarity between the query and the documents in the database in their lower dimensional representation results in the closest document matching to be retrieved. The similarity measure commonly used is the cosine measure.

The idiosyncratic patterns in speech are likely to be speaker dependent. These patterns are noticed more in unrestricted conversational speech. They are not so pronounced in read speech or news bulletin type of speech. In this paper the idea of authorship attribution and semantic retrieval are combined to perform speaker recognition on the NIST 2003 extended data task [12]. The paper is organised as follows. The next section describes the theory of latent semantic analysis. Section 3 describes the database used for the task. The experiments and the results arrived at are reported in Section 4 and 5 respectively. Section 6 summarises the study.

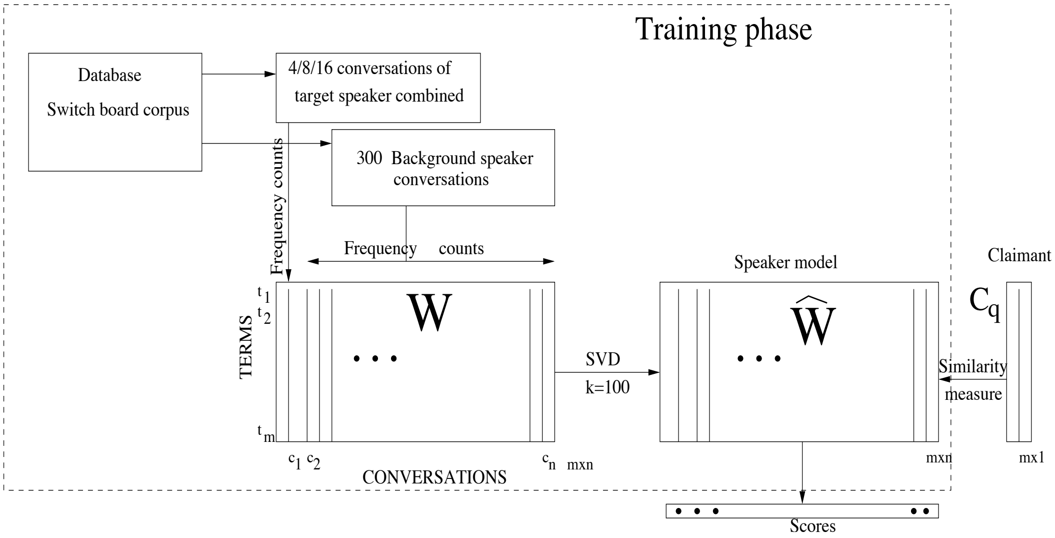

Figure 1: Block diagram of the proposed system

2. Latent semantic analysis

For every conversation of a speaker, the co-occurrences of words and phrases (n-grams) called terms, are sought to be captured in the latent semantic analysis frame work. Latent semantic analysis as described in [11] [13] uses singular value decomposition, a technique closely related to eigen value decomposition. A *term* × *conversation* matrix W is constructed with rows representing terms and columns representing conversations. Each element of W contains the frequency count of term i in conversation j. Figure 1 depicts the the block diagram of proposed system. From the switch board corpus a set of 4/8/16 conversations of the target speaker are considered and the frequency counts of the terms in a conversation are used to construct a column of the W matrix. To facilitate decision making, conversations of background speakers are chosen as additional columns of the W matrix. For each conversation by a background speaker, one column of W is derived. The similarity between a test conversation and the model reveals that the conversation is either similar to that of the target speaker or background speakers. The use of background speakers also helps in test score normalisation. The matrix W, is decomposed using SVD into a product of three matrices such that $\hat{W} = T.S.C^T$, where T and C have orthonormal columns, S is diagonal and $\hat{W}$ is the reconstructed matrix. We retain only k singular values of S and the corresponding columns of T and C. The resulting matrix $\hat{W} = T_k S_k C_k^T$ is the closest matrix of rank k to W in the least square sense. These k linearly independent components capture the major structural association in the data and remove the noise. One model $(\hat{W})$ is derived for each set of conversations of the target speaker. Term usage by a speaker is directly used in building the model since this could be an indicator of speaker idiosyncrasies, as in authorship attribution studies. The term counts in a conversation, can be weighted by the inverse speaker entropy of the term in the corpus [9]. The entropy of term t is given by

$$\begin{aligned} E_t &= -\sum_i P_t(s_i)\log P_t(s_i) \\ \text{where} \quad P_t(s_i) &= \frac{N_t(s_i)}{N_t(C)} \end{aligned}$$

$P_t(s_i)$ is the fraction of terms spoken by speaker s_i in the entire set of conversations considered for that particular model, including conversation from background speakers. This weighing emphasises speaker specific terms.

In testing the system, a vector of terms in the claimant speaker's conversation (X_q) is constructed in a manner similar to a column of the W matrix. The representation of X_q in the reduced k dimensional space is given by $C_q = X_q^T T S^{-1}$. The similarity between the claimant speaker (C_q) and the target model (appropriate column of $\hat{W}$) is computed. Depending on the threshold the claim is accepted or rejected. Different similarity measures may be used in arriving at the scores.

3. Database

As part of the NIST 2003 [12] speaker evaluation plan, the extended data task involves speech of 5 minutes duration per conversation for each speaker. This data is part of the switchboard corpus phase 2 and 3. Set (I) of the extended data task was considered for our study. This contains 31 speakers. For each speaker different sets of 4, 8 or 16 conversations involving him/her were used to build models for that speaker. A total of 265 models were thus built as dictated by the evaluation. The claimant utterances were of 2 minutes duration. A total of 3,663, tests were conducted against these 265 models. Auxiliary information in terms of transcription of the entire corpus (word error rate $\simeq 40$), pitch contour estimates for the entire database, GMM scores and bigram speaker model scores for the test utterance were made available along with the acoustic signal by different institutions as part of the evaluation.

4. Experiments

All n-grams of order 1 to 5 occurring in the conversations of the target speaker and background speakers that form the columns of the W matrix for the target model are potential terms. An arbitrary cutoff on the frequency counts of the n-grams is used to limit the number of terms. For the best performing experiment we had about 5,000 to 7,500 terms. In different experiments we used 30 or 300 background speakers. Thus we had a W matrix of size $terms \times 31$ or $terms \times 301$. The order of decomposition in SVD depends on the balance between minimising the reconstruction error and maximising the noise suppression. This results in an order of 17 or 34 for models with 30 or 300 background speakers respectively. The performance of these systems were suboptimal and are not reported here. Another factor in deciding on the order of decomposition is the number of abstract semantic concepts present in the set of conversations being modelled. Based on this , use of order 100 seems to perform optimally for the current task. The representation of the claimant conversation in the lower dimensional SVD space is obtained. The similarity between this claimant and the target model (first column of $\hat{W}$) results in a score for that model. The scores obtained between the claimant model and the background speakers (rest of the columns of $\hat{W}$) is used for test utterance normalisation. The similarity measures used are cosine measure, Pearson correlation and Jaccard similarity as defined below:

Given two vectors $\mathbf{x}_a, \mathbf{x}_b$ the cosine measure is given by:

$$S^{(c)}(\mathbf{x}_a, \mathbf{x}_b) \quad = \quad \frac{\mathbf{x}_a^T \mathbf{x}_b}{||\mathbf{x}_a||_2 ||\mathbf{x}_b||_2} \qquad (1)$$

The normalised Pearson correlation is defined as:

$$S^{(p)}(\mathbf{x}_a, \mathbf{x}_b) \quad = \quad \frac{1}{2} \frac{(\mathbf{x}_a - \bar{\mathbf{x}}_a)^T (\mathbf{x}_a - \bar{\mathbf{x}_b})}{||(\mathbf{x}_a - \bar{\mathbf{x}}_a)||_2 ||(\mathbf{x}_a - \bar{\mathbf{x}_b})||_2} \quad (2)$$

where $\bar{\mathbf{x}}_a$ is the average value of $\mathbf{x}$ over all dimensions and $||\mathbf{x}||_2$ denotes the L_2 norm of $\mathbf{x}$.

The binary Jaccard coefficient measures the degree of overlap between two sets, and is computed as the ratio of shared attributes (words) of $\mathbf{x}_a$ and $\mathbf{x}_b$ to the number possessed by $\mathbf{x}_a$ or $\mathbf{x}_b$. Extending this to discrete non-negative features the similarity is given by:

$$S^{(j)}(\mathbf{x}_a, \mathbf{x}_b) \quad = \quad \frac{\mathbf{x}_a^T \mathbf{x}_b}{||\mathbf{x}_a||_2 ||\mathbf{x}_b||_2 - \mathbf{x}_a^T \mathbf{x}_b} \qquad (3)$$

Table 1: Decision cost factor for different term types with and without weighing by entropy for the three similarity measures

Term type	Similarity measure	With entropy weighing	No entropy weighing
Unigrams	Correlation	24.32	21.35
	Cosine	24.27	21.74
	Jaccard	23.32	20.89
Bigrams	Correlation	49.31	30.99
	Cosine	48.43	32.62
	Jaccard	49.98	31.00
2-5 grams	Correlation	28.12	32.34
	Cosine	30.39	32.46
	Jaccard	29.11	31.07
1-5 grams	Correlation	19.34	19.09
	Cosine	19.68	19.66
	Jaccard	19.02	19.02

5. Results

In order to arrive at the most optimal performance, different combinations of terms were tried. The effectiveness of usage of words (unigrams) as terms as in authorship attribution studies was carried out. The performance of bigrams alone, combination of all n-grams except unigrams and higher order n-grams, were explored. The results are tabulated in Table 1 in terms of the optimal decision cost factor for different similarity measures. We observed that the best system was obtained when we used all the n-grams of order 1 to 5 in the term list, and an SVD order of decomposition of 100. The performance of all the similarity measures is about the same. Unlike the case in information retrieval in which weighing with inverse entropy gives more weightage to content words (infrequent words) weighing does not help in speaker recognition. In speaker recognition the frequent use of function words may be a better indicator of speaker idiosyncrasies. This is reflected in the performance of the systems in column (4) of Table 1.

The DET plots for the system using n-grams of order 1 to 5 and SVD dimension 100 is shown in Figure 2. The DET plots corresponding to the same data set for the available auxiliary information are also plotted. We observe that the equal error rate (EER) for the LSA system and the language model (LM) based auxiliary scores is about the same. Another system which gives speaker recognition scores based on AANN models trained and tested on the acoustic features extracted only from 10 frequently occurring word segments in the database is also plotted. Figure 3 shows the combination of the scores from different systems using the sum rule [14]. We notice that the addition of LSA scores to any of the other systems improves the overall performance. The best per-

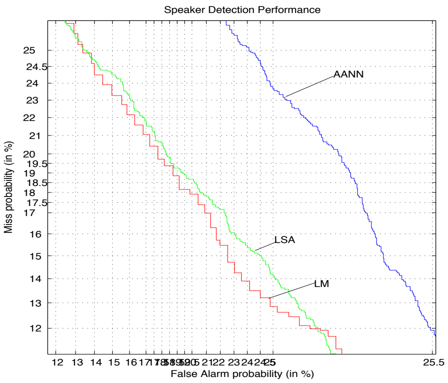

Figure 2: EER plot for individual systems

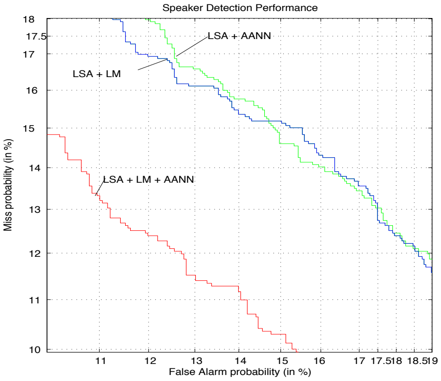

Figure 3: EER plot for combined systems

formance is obtained by combining scores from all three systems. This suggests that there is significant complimentary information that can be derived from the LSA system which would help improve the speaker recognition performance. It is observed that combining these scores with the auxiliary GMM scores further reduces the EER.

6. Summary

In this paper we have presented an approach to speaker recognition which is based on the principles derived from authorship attribution studies, idiolectic speaker recognition and latent semantic analysis. It is shown that the performance using text transcripts of low quality ($\simeq 40\%$ WER) we obtain performance similar to other systems. The advantage of the proposed system is that the information seems to be of complimentary nature, which can be exploited to improve the overall performance.

7. References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[2] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459–469, Apr. 2002.

[3] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 1687–1697, 1972.

[4] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Orlando, Florida, USA, May 2002, vol. 1, pp. 141–144.

[5] F. Mosteller and D. Wallace, *Inference and Disputed Authorship*, Addison-Wesley, Reading, Mass., 1964.

[6] R. A. Bosch and J. A. Smith, "Separating hyperplanes and the authorship of the disputed federalist papers," *American Mathematical Monthly*, vol. 7, no. 105, pp. 601–608, Aug. 1998.

[7] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic authorship attribution," in *Proc. nineth Conf. European Chap. Assoc. Computational Linguistics*, Bergen, Norway, Jun. 1999, pp. 158–164.

[8] Glenn Fung, "The disputed federalist papers: SVM and feature selection via concave minimization," in *Proc. 2003 Conf. Diversity in Computing, TAPIA'03,*, Atlanta, Georgia, Oct. 2003, pp. 42–46.

[9] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. EUROSPEECH* , Scandinavia, Sept. 2001, vol. 4, pp. 2521–2524.

[10] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[11] S. C. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Information*, vol. 41, no. 6, pp. 391–407, 1990.

[12] The 2003 NIST Speaker Recognition Evaluation, "http://www.nist.gov/speech/tests/spk/2003/," .

[13] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process*, vol. 25, pp. 259–284, 1998.

[14] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, Mar. 1998.