TIME-ALIGNED SVD ANALYSIS FOR SPEAKER IDENTIFICATION

Patrick J. Clemins, Heather E. Ewalt, and Michael T. Johnson

Speech and Signal Processing Lab
Marquette University, Milwaukee, WI
patrick.clemins@mu.edu, heather.ewalt@mu.edu, and mike.johnson@mu.edu

ABSTRACT

This paper presents a time-aligned singular value decomposition (SVD) analysis for speaker identification. SVD analysis has been used for fast spectral matching based on a global representation of an entire utterance. We incorporate temporal normalization directly into the decomposition by using a dynamic time warping (DTW) path to time-align the rows of the feature matrix prior to SVD analysis. Speaker identification results using the TI-46 database indicates that the time-aligned SVD significantly improves accuracy for most threshold choices.

1. INTRODUCTION

Applications of speaker identification for voice authentication or other security purposes require computationally efficient algorithms that are able to build accurate speaker models with limited enrollment data, sometimes with as little as a single utterance [1]. An example of such an application is the use of verbal passphrases for computer login and re-entry. The newest versions of Macintosh's operating system, MacOS, have implemented a global utterance matching algorithm based on a Singular Value Decomposition (SVD) of each utterance [2]. This approach, based on a subspace interpretation of an utterance's feature space [3], attempts to separate the temporal and spectral information as generated by a specific speaker.

Advantages of this approach include a very fast and easily implemented identification metric as well as an ability to deal with extremely limited amounts of training data. Other methods such as Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs), which need a relatively large amount of data for statistical parameter estimation, are unsuited to this task.

The method used for the MacOS application was combined with a DTW algorithm to improve overall identification accuracy. The results indicated that the DTW analysis provided a much better separation between imposter and target speaker scores than the SVD scoring. The SVD scores in isolation yielded only mediocre accuracy and have significant potential for improvement. However, since the types of errors made by the SVD metric appeared to be somewhat orthogonal to those of the DTW algorithm, their inclusion was able to make a measurable difference in overall system accuracy.

This original method used SVD and DTW independently of each other, using a simple two-dimensional threshold as a decision criterion. In particular, direct SVD analysis does not take advantage of the temporal information generated by a DTW alignment. To do this, we propose a new SVD-based method using a time-warped feature matrix from the test utterance, so that the SVD analysis can be re-aligned to take advantage of the temporal information generated by the DTW analysis. The hypothesis is that this extra information will improve separation between the target speaker scores and the imposter scores.

2. SVD ANALYSIS

Each utterance is represented by an M x N matrix of frames, with rows containing the feature information for a frame and columns containing a specific feature over time. For this experiment, twelve LPC-derived cepstral coefficients and an energy measure are used as the primary features. In addition, deltas and delta deltas are also computed for each of these features, creating a feature vector of 39 values. Frames are 10ms with no overlap, giving typical feature matrices of about 100 x 39.

The resulting feature matrix, F, can be represented as an orthonormal decomposition [4]:

$$F = U S V^T. \quad (1)$$

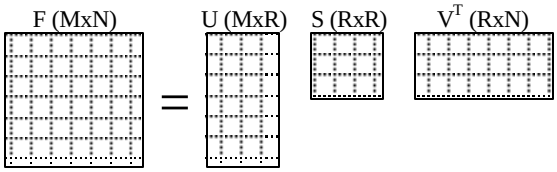

Figure 1 - A Representation of SVD

U is an M x R left singular matrix, S is an R x R diagonal matrix of singular values, and V is an N x R right singular matrix. Both U and V are column-orthonormal so $U^TU = V^TV = I$. The R singular values of S are in non-increasing order, so the full decomposition can easily be reduced to a 'filtered' version of the original matrix by decreasing the number of singular values used. This process is generally referred to as reduced singular value decomposition, and the order R is determined by either arbitrarily fixing the number of desired singular values or by taking all which are greater than a given threshold. As discussed in [2], there are multiple interpretations of the role of the matrices U and V with respect to the feature vector created by a speech utterance. One reasonable viewpoint is that U creates an orthogonal projection of the spectral features and that V creates an orthogonal projection of the temporal content.

Following the derivation given in [2], we can use this viewpoint to create a distance metric to evaluate closeness between two decompositions. To evaluate the spectral characteristic's similarity between two utterances, a SVD is performed on the reference template, F_k, and the test utterance, F_l:

$$F_k = U_k S_k V_k^T, \quad (2)$$
$$F_l = U_l S_l V_l^T. \quad (3)$$

Using the V matrices to map the test utterance singular value matrix, S_l, onto the reference utterance subspace, the following metric can be established [2]:

$$D_{l|k} = (V_l^T V_k)^T S_l (V_l^T V_k). \quad (4)$$

The degree to which $D_{l|k}$ deviates from a diagonal matrix is related to the amount that the test utterance differs from the reference utterance. As V_t tends to V_k, $D_{t|k}$ becomes a diagonal matrix, converging to $S_t = S_k$. One way to measure the diagonality of the $D_{j|S'}$ matrix is by taking the Frobenius norm of the off-diagonal terms:

$$|D_{l|k}|_F = \sum_i \sum_j d_{ij}^2, \; i \neq j. \quad (5)$$

The closer this measure is to zero, the more similar the two utterances.

3. TIME-ALIGNED SVD

Dynamic time warping (DTW) is a well-known algorithm for finding a frame-by-frame mapping from a test utterance to a reference template. This mapping is done to eliminate temporal variation between one utterance and the next. By normalizing the temporal variation, the characteristics of the test and reference utterances can be compared directly using simple spectral difference measures.

The DTW algorithm used here implements the standard dynamic programming version of the algorithm, using typical path constraints [5]. Mahalanobis distance with a diagonal covariance matrix is used as the distance measure. To create the speaker template, the median-length training utterance is used for template initialization. Time-warped training utterances are averaged to create a new template.

To incorporate the temporal information gained from the DTW time-alignment directly into the SVD method as desired, the warping path is used to map the original feature matrix into a new time-normalized feature matrix prior to the decomposition process. The time aligned feature matrix, F_t: is decomposed as before:

$$F_t = U_t S_t V_t^T. \quad (6)$$

The matrix resulting from the mapping of S_t onto the reference utterance subspace is given as $D_{t|k}$:

$$D_{t|k} = (V_t^T V_k)^T S_t (V_t^T V_k). \quad (7)$$

The Frobenius norm of the off-diagonal terms is then calculated to generate a SVD score for the test utterance. To compare the results and measure improvement, the SVD method is done on both the original and time-aligned feature matrices.

4. RESULTS

4.1. Time-aligned SVD vs. Standard SVD

The TI 46-Word Speaker-Dependent Isolated Word Corpus (TI46) was used to conduct the analysis of the new time-aligned SVD method. The full corpus consists of eight male and eight female speakers repeating short words and letters. The utterances are sampled at 12.5 kHz and are about one second in length. There are ten training

Word	Speakers associated with a word			
rubout	m2	m3	m5	m6
enter	m3	m4	m7	m8
erase	m1	m2	m6	m7
repeat	m1	m4	m5	m8

Table 1 – Table of Words Spoken by Each Speaker

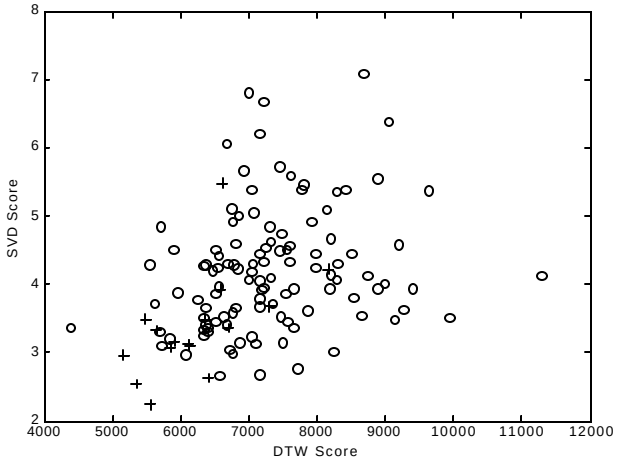

Figure 2 – SVD Without Time-Alignment
'+' = Target Speakers
'o' = Imposters

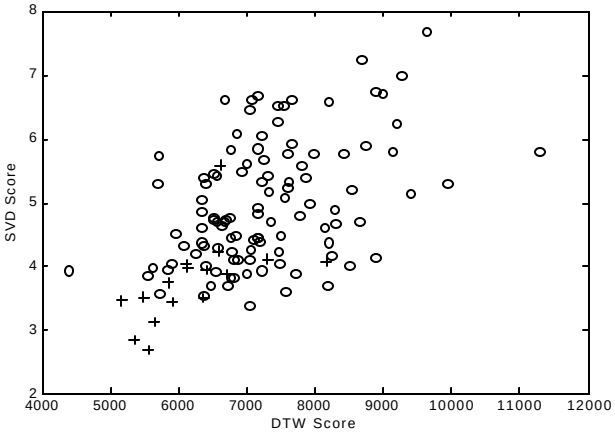

Figure 3 – SVD With Time-Alignment
'+' = Target Speakers
'o' = Imposters

utterances and 16 test utterances for each word and letter. Because of the relatively short duration of the utterances in this corpus, it was unknown how much influence the addition of temporal normalization to the SVD algorithm would have; however, since passwords and passphrases are often fairly short, this difficulty is appropriate for the task in mind.

Both the unaligned and aligned SVD methods were used to identify the males in the TI-46 corpus. The females' utterances were not used since the gender difference would make the identification task easier and cause the percentage of false positives to artificially decrease. By using one of the eight speakers as a target and the remaining seven as imposters, a set of eight target tests and 56 imposter tests was created for each speaker. With each male speaker as a target for two different words, a complete experimental set of 128 target tests and 896 imposter tests was constructed. Table 1 shows a table of which words are used for which speaker. Only the two syllable words in the corpus were chosen for these experiments. An SVD score was generated for both the time-aligned matrices and original feature matrices for each of the 1024 tests. The size of the singular value matrix S was fixed at 10 for this portion of the experiment. An example plot of the scores for male number three for both methods is shown in figures 2 and 3. Notice that the time-aligned SVD method moves the target scores lower in the vertical direction in relation to the imposter scores.

A Receiver Operating Characteristics (ROC) curve showing false accept versus false reject rates was also generated for each method. This plot for a singular value matrix of order 10 is shown in figure 4. The new method dominates for all thresholds. Note that these error rates do not include the influence of DTW scores, which would improve both methods. The equal error rate for the original SVD method is 33.4%, while the equal error rate for the aligned SVD method is 28.2%.

4.2. Number of Singular Values vs. EER

The second experiment performed was to evaluate performance of both algorithms as a function of the order of the SVD used. The number of singular values R was varied between 2 and 39 and the equal error rate (EER) calculated for each trial. A plot of EER vs. R for both methods is shown in Figure 5. From the plot it can be shown that the time-aligned SVD scores provide greater separation between target scores and imposter scores for smaller singular value matrices. The two error rates converge as the order of the singular value matrix approaches the number of features.

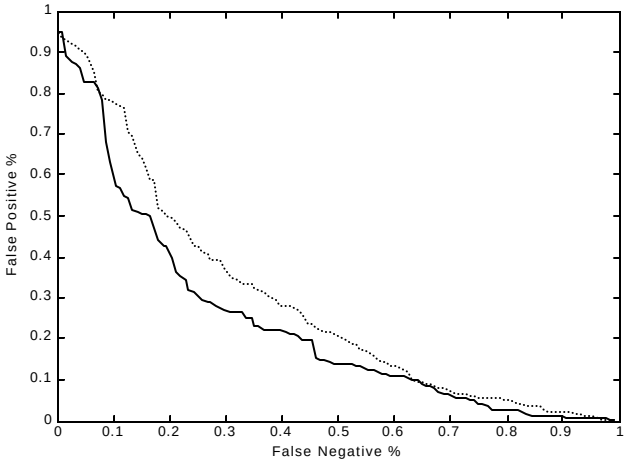

Figure 4 – ROC Curve for SVD alone
Solid Line = Aligned SVD
Dotted Line = Original SVD

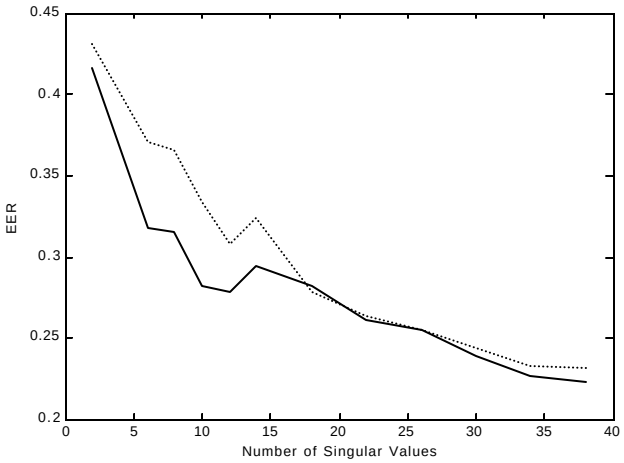

Figure 5 – EER vs. Number of Singular Values Used
Solid Line = Aligned SVD
Dotted Line = Original SVD

5. CONCLUSIONS

The use of the alignment path obtained through DTW time normalizes the feature matrix to enable direct rather than indirect incorporation of temporal information into the SVD scoring metric. The results indicate that this new metric achieves a greater separation between the target scores and imposter scores. Using 10 singular values in the SVD decomposition, an improvement was made in EER from 33.4% to 28.2% using only the SVD score. Overall, the time-aligned SVD approach may improve system accuracy for tasks where enrollment data is severely limited.

6. REFERENCES

[1] Campbell, J.P. Jr., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, Vol. 85, pp. 1437-1462, Sept. 1997.

[2] Bellegarda, Jerome R., Devang Naik, Matthias Neeracher, and Kim E.A. Silverman, "Language-Independent, Short-Enrollment Voice Verification Over A Far-Field Microphone," *Proc. 2001 ICASSP*, Salt Lake City, UT, pp. 145-148, May 2001.

[3] Y. Ariki and K. Doi, "Speaker Recognition Based on Subspace Methods," *Proc. ICSLP*, Tokohama, Japan, pp. 1859-1862, September 1994.

[4] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins, Baltimore, MD, Second Ed., 1989.

[5] Sakoe, H., and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 26, pp. 43-49, Feb. 1978.