LANGUAGE–INDEPENDENT, SHORT–ENROLLMENT VOICE VERIFICATION OVER A FAR–FIELD MICROPHONE

Jerome R. Bellegarda, Devang Naik, Matthias Neeracher, and *Kim E.A. Silverman*

Spoken Language Group
Apple Computer, Inc.
Cupertino, California 95014, USA

ABSTRACT

A new approach is presented for the dual verification of speaker identity and verbal content in a text-dependent voice authentication system. The application considered is desktop voice login over a far-field microphone. Each speaker is allowed to select his or her own keyphrase, and enrollment is limited to four instances of the keyphrase, each 1 to 2 seconds of speech. The approach decouples the analysis of speaker and verbal content information, so as to use two light-weight components for verification: a spectral matching component based on a global representation of the entire utterance, and a temporal alignment component based on more conventional frame-level information. The resulting integration is language-independent, and experiments with deliberate imposture show an equal error rate figure of approximately 4%. The approach has been commercially deployed in the "VoicePrint Password" feature of MacOS 9.

1. INTRODUCTION

Voice authentication refers to the process of accepting or rejecting the identity claim of a speaker on the basis of individual information present in the speech waveform [1]. It has received increasing attention over the past two decades, as a convenient, user-friendly way of replacing (or supplementing) standard password-type matching [2]. In this paper, we will focus on a desktop login application, where access to a personal computer can be granted or denied on the basis of the user's identity. In that context, the authentication system must be kept as unintrusive as possible, which typically entails the use of a far-field microphone (e.g., mounted on the monitor) and a very small amount of enrollment data (5 to 10 seconds of speech).

The scenario of choice for desktop voice login is text-dependent verification, which is logistically closest to that of a typed password. Each registered speaker is asked to select a keyphrase of his or her own choosing, and then use that keyphrase in both training and recognition trials. Assuming the user maintains the confidentiality of the keyphrase, this offers the possibility of verifying the spoken keyphrase in addition to the speaker identity, thus resulting in an additional layer of security. Hence the interest in authentication methods which can concurrently verify both speaker characteristics and verbal content (cf. [3]).

This scenario entails the comparison of the acoustic sequence uttered during recognition (verification utterance) with the aggregated acoustic evidence collected during training (keyphrase-specific reference speaker model). This is typically done using HMM technology with Gaussian mixture distributions (see, e.g., [4], [5]). Conceptually, the verification utterance is aligned against the appropriate subword HMMs constructed from the relevant reference speaker model, and the likelihood of the input speech matching the reference model is calculated. If this likelihood is high enough, the speaker is accepted as claimed. This approach faces scarce data problems, however, when enrollment is severely limited, as is the case here. Variance estimation is of particular concern, as the underlying Gaussian mixture distributions run the risk of being too sharp and overfitting the training data [6].

This paper proposes an alternative solution, loosely based on a divide and conquer strategy. Rather than using a single paradigm to verify both speaker and verbal content simultaneously, we attempt to decouple the two and adopt a different (light-weight) algorithm for each of them: one primarily for global spectral content matching, and the other mostly for local temporal alignment.

For global spectral content matching, we rely on an utterance-level representation obtained by integrating out frame-level information through the use of singular value decomposition (SVD). With this new representation, it is possible to relate the verification evidence and reference model through a simple linear transformation. Then the speaker verification problem becomes a matter of computing the appropriate distance between these two entities. To this end, we derive and adopt a new metric which arises naturally from the SVD framework.

For temporal alignment, we use simple dynamic time-warping (DTW). Although HMMs can more efficiently model statistical variation in spectral features, here DTW is sufficient, because the SVD approach already takes care of spectral matching, and therefore the requirements on DTW are less stringent than is usually the case. The overall system, combining both SVD and DTW components, is language-independent. Each component computes a separate likelihood score for how well the input speech matches the reference model. The accept/reject decision is then based on the combination of these two scores.

The paper is organized as follows. The next section briefly reviews feature extraction, which is common to the two components. Section 3 describes the SVD framework and associated metric. In Section 4, we discuss integration with the DTW component. Finally, Section 5 reports experimental results illustrating the resulting benefits.

2. FEATURE EXTRACTION

We extract spectral feature vectors every 10ms, using short-term FFT followed by filter bank analysis to ensure a smooth spectral envelope. (This is important to provide a stable representation from one repetition to another of a particular speaker's utterance.) To represent the spectral dynamics, we also extract, for every frame, the delta and delta-delta parameters. After concatenation, we therefore end up with a sequence of M feature vectors (frames) of dimension N, where, for a typical utterance, $M \approx 200$ and $N \approx 40$. This sequence is the input to both the SVD component and the DTW component of the proposed method.

3. SVD FRAMEWORK

From the above, each utterance is represented by a $M \times N$ matrix of frames, say F, where each row represents the spectral information for a frame and each column represents a particular spectral band over time. We can therefore compute the SVD of the matrix F, as [7]:

$$F = U\,S\,V^T\,, \tag{1}$$

where U is the $(M \times R)$ left singular matrix, S is the $(R \times R)$ diagonal matrix of singular values, V is the $(N \times R)$ right singular matrix, $R < \min(M, N)$ is the order of the decomposition, and ^{T} denotes matrix transposition. As is well known, both U and V are column-orthonormal, i.e., $U^T U = V^T V = I_R$, the identity matrix of order R. For reasons to become clear shortly, we refer to (1) as the decomposition of the utterance into *single-utterance* singular elements U, S, and V.

Such whole utterance representation has been considered before: see, e.g., [8]. The resulting parameterization can be loosely interpreted as conceptually analogous to the Gaussian mixture parameterization in the HMM framework. The main difference is that the Gaussian mixture approach is implicitly based on a sub-word unit (such as a phoneme), whereas the SVD approach operates on the entire utterance, which introduces more smoothing.

It is intuitively reasonable to postulate that some of the singular elements will reflect more speaker information and some others more verbal content information. But it is not completely clear exactly which reflects what. In [8], a case was made that speaker information is mostly contained in V. Speaker verification was then performed using the Euclidean distance after projection onto the "speaker subspace" defined by V, on the theory that in that subspace utterances from the true speaker have greater measure. This is illustrated in Fig. 1, top figure. In that interpretation, each row of V^T can be thought of as a basis vector spanning the global spectral content of the utterance, and each row of US represents the degree to which each basis vector contributes to the corresponding frame.

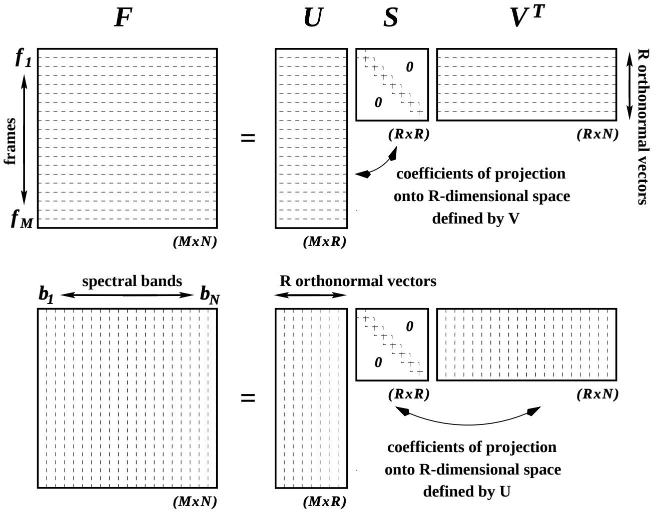

Fig. 1. Two Equivalent Views of SVD Decomposition.

But an equally compelling case could be made under the (dual) assumption that verbal content information is mostly contained in U. In this situation speaker verification could conceivably be performed after projection onto the "content subspace" spanned by U. One would simply compute distances between reference and verification utterances in that subspace, on the theory that a large distance between two utterances with the same verbal content would have to be attributed to a speaker mismatch. This is illustrated in Fig. 1, bottom figure. In that interpretation, each column of U is a basis vector spanning the verbal content of the utterance, and each column of SV^T represents the degree to which each basis vector contributes to the corresponding spectral band.

So, what is the "correct" way to make use of the representation (1)? Arguably, the problem is not so much in the precise interpretation of the singular elements as it is in the specification of a distance measure perhaps more appropriate than the Euclidean (or Gaussian) distance to evaluate closeness in this representation. The following justifies and adopts a new metric specifically tailored to the SVD framework.

Assume, without loss of generality, that (1) is associated with a particular training utterance, say the jth utterance, from a given speaker, and consider the set of all training utterances from that speaker. This set will be represented by a $\tilde{M} \times N$ matrix, with $\tilde{M} \approx JM$, where J is the number of training utterances for the speaker. Denoting this $\tilde{M} \times N$ matrix by $\tilde{F}$, it can be decomposed as:

$$\tilde{F} = \tilde{U}\,\tilde{S}\,\tilde{V}^T\,, \tag{2}$$

with analogous definitions and properties as in (1). Obviously, the set of all training utterances contains the jth utterance, so by selecting the appropriate M rows of $\tilde{F}$, we can define:

$$\tilde{F}_j = F = \tilde{U}_j\,\tilde{S}\,\tilde{V}^T\,, \tag{3}$$

as the decomposition of the jth utterance into *multiple-*

utterance singular elements $\tilde{U}_j$, $\tilde{S}$, and $\tilde{V}$. Presumably, from the increased amount of training data, the matrices $\tilde{S}$ and $\tilde{V}$ are somewhat more robust versions of S and V, while $\tilde{U}_j$ relates this more reliable representation (including any embedded speaker information) to the original jth utterance. The equality:

$$\tilde{U}_j\, \tilde{S}\, \tilde{V}^T = U\, S\, V^T\,, \qquad (4)$$

follows from (1) and (3). To cast this equation into a more useful form, we now make use of the (easily shown) fact that the matrix $(V^T\tilde{V})$ is (both row- and column-) orthornormal. After some algebraic manipulations, we eventually arrive at the expression:

$$\tilde{S}\, (\tilde{U}_j^T\, \tilde{U}_j)\, \tilde{S} = (V^T\, \tilde{V})^T\, S^2\, (V^T\, \tilde{V})\,. \qquad (5)$$

Since both sides of (5) are symmetric and positive definite, there exists a $(R \times R)$ matrix $D_{j|\tilde{S}}$ such that:

$$D^2_{j|\tilde{S}} = \tilde{S}\, (\tilde{U}_j^T\, \tilde{U}_j)\, \tilde{S}\,. \qquad (6)$$

Note that, while $\tilde{U}^T\tilde{U} = I_R$, in general $\tilde{U}_j^T\tilde{U}_j \neq I_R$. Thus $D^2_{j|\tilde{S}}$ is closely related, but not equal, to $\tilde{S}^2$. Only as the single-utterance decomposition becomes more and more consistent with the multiple-utterance decomposition does $D^2_{j|\tilde{S}}$ converge to $\tilde{S}^2$.

Taking (6) into account and again invoking the orthonormality of $(V^T\,\tilde{V})$, the equation (5) is seen to admit the solution:

$$D_{j|\tilde{S}} = (V^T\, \tilde{V})^T\, S\, (V^T\, \tilde{V})\,. \qquad (7)$$

Thus, the orthornormal matrix $(V^T\,\tilde{V})$ can be interpreted as the rotation necessary to map the single-utterance singular value matrix obtained in (1) onto (an appropriately transformed version of) the multiple-utterance singular value matrix obtained in (2). Clearly, as V tends to $\tilde{V}$ (meaning U also tends to $\tilde{U}_j$) the two sides of (7) become closer and closer to a diagonal matrix, ultimately converging to $S = \tilde{S}$.

This suggests the following metric to evaluate how well a particular utterance j is consistent with the (multiple-utterance) speaker model: compute the quantity $D_{j|\tilde{S}} = (V^T\tilde{V})^T S\, (V^T\tilde{V})$, per (7), and measure how much it deviates from a diagonal matrix. For example, one way to measure the deviation from diagonality is to calculate the Frobenius norm of the off-diagonal elements of the matrix $D_{j|\tilde{S}}$.

This further suggests an alternative metric to evaluate how well a verification utterance, uttered by a speaker ℓ, is consistent with the (multiple-utterance) model for speaker k. Indexing the single-utterance elements by ℓ, and the multiple-utterance elements by k, we define:

$$D_{\ell|k} = (V_\ell^T\, V_k)^T\, S_\ell\, (V_\ell^T\, V_k)\,, \qquad (8)$$

and again measure the deviation from diagonality of $D_{\ell|k}$ by calculating the Frobenius norm of its off-diagonal elements. By the same reasoning as before, in this expression the matrix $(V_\ell^T\, V_k)$ underscores the rotation necessary to map S_ℓ onto (an appropriately transformed version of) S_k. When V_ℓ tends to V_k, $D_{\ell|k}$ tends to S_k, and the Frobenius norm tends to zero. Thus, the deviation from diagonality can be expected to be less when the verification utterance comes from speaker $\ell = k$ then when it comes from a speaker $\ell \neq k$. Clearly, this distance measure is better tailored to the SVD framework than the usual Euclidean (or Gaussian) distance. It can be verified experimentally that it also achieves better performance.

The SVD component thus operates as follows. During enrollment, each speaker $1 \leq k \leq K$ to be registered provides a small number J of training sentences. For each speaker, the enrollment data is processed as in (2), to obtain the appropriate right singular matrix V_k. During recognition, the input utterance is processed as in (1), producing the quantity S_ℓ and V_ℓ. Then $D_{\ell|k}$ is computed as in (8), and the deviation from diagonality is calculated. If this measure falls within a given threshold, then the speaker is accepted as claimed. Otherwise, it is rejected.

4. INTEGRATION WITH DTW

The SVD approach clearly does not take full advantage of temporal information, since it integrates out frame-level information. Because of the linear mapping inherent in the decomposition, it is likely that the singular elements only encapsulate coarse time variations, and smooth out finer behavior. Unfortunately, detecting subtle differences in delivery is often crucial to thwarting non-casual impostors, who might use their knowledge of the true user's speech characteristics to deliberately mimic his or her spectral content. Thus, a more explicit temporal verification should be added to the SVD component to increase the level of security against such determined impersonators.

We propose a simple DTW approach for this purpose. Although HMM techniques have generally proven superior, in the present case the SVD approach already contributes to spectral matching, so the requirements on any supplementary technique are less severe. As it turns out, DTW suffices, in conjunction with the SVD component, to carry out verbal content verification in our application.

The DTW component implements the classical dynamic time warping algorithm (cf., e.g., [9]). During training, the J training utterances provided by each speaker are "averaged" to define a representative reference utterance u_R. This is done by setting the length of u_R to the average length of all J training utterances, and warping each frame appropriately to come up with the reference frame at that time. During verification, the input sequence of M feature vectors of dimension N, say u_V, is acquired and compared to the reference model u_R. This is done by aligning the the time axes of u_V and u_R, and computing the degree of similarity between them, accumulated from the beginning to the end of the utterance on a frame by frame basis. Various distance measures are adequate to perform this step, including the usual Gaussian distance. If the degree of similarity is high enough, the speaker is accepted as claimed. Otherwise, it is rejected.

For each verification utterance, two scores are produced:

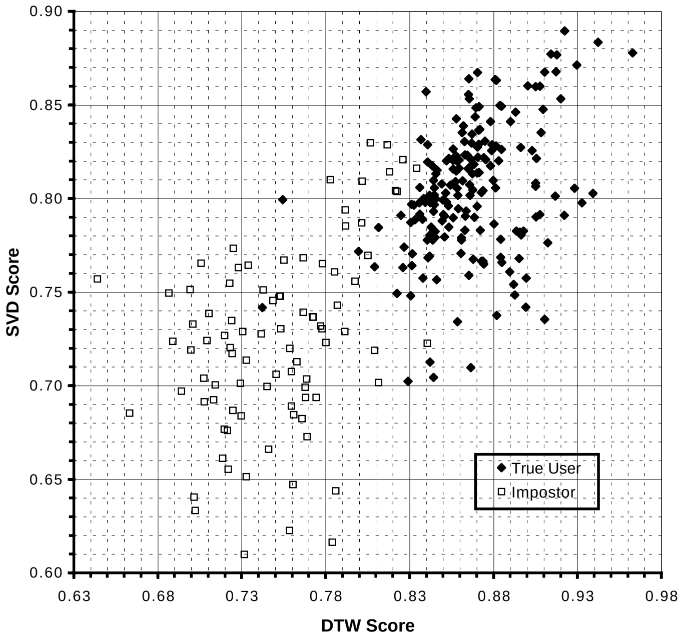

Fig. 2. Performance Space of SVD+DTW Approach.

the deviation from diagonality score from the SVD component, and the degree of similarity from the DTW component. There are therefore several possibilities to combine the two components. For example, it is possible to combine the two scores into a single one and base the accept/reject decision on that single score. Alternatively, one can reach a separate accept/reject decision for each component and use a voting scheme to form the final decision.

For simplicity, we opted for the latter. Thus, no attempt is made to introduce conditional behavior in one component which depends on the direction taken by the other. The speaker is simply accepted as claimed only if both likelihood scores are high enough.

5. EXPERIMENTAL RESULTS

Experiments were conducted using a set of 93 speakers, $K = 48$ true users and $K' = 45$ impostors. True users enrolled by speaking their keyphrase $J = 4$ times. They also provided four instances of a voice login attempt, collected on different days. This resulted in a total of 191 true test utterances, across which the minimum, average, and maximum sentence length were 1.2, 1.8, and 3 seconds, respectively.

To increase the severity of the test, each impostor was dedicated to a particular speaker, and was selected on the basis of his/her apparent "closeness" to that user, as reflected in his/her speech characteristics. For example, to impersonate a male speaker who grew up in Australia, we chose another male speaker with an Australian accent. Further, each impostor was given access to the original enrollment keyphrases from the true speaker, and was encouraged to mimic delivery as best as s/he could. This was to reflect the high likelihood of deliberate imposture in desktop voice login, where the true user is typically known to the impostor. (On the other hand, given this application and in view of Apple's target market, we deemed unnecessary to consider more sophisticated attempts like technical imposture [10].) Each impostor provided two distinct attempts, for a total of 90 impostor test utterances.

The results are plotted in Fig. 2. For the appropriate combination of thresholds, the above system leads to 0 false acceptances and 20 false rejections (10.4%). After tuning to obtain an equal number of false acceptances and false rejections, we observed approximately a 4% equal error rate.

6. CONCLUSIONS

We have presented a novel approach to the dual verification of speaker identity and verbal content in a desktop voice login application. Because enrollment is limited to an average of about 7 seconds of speech, usual HMM-based methods are prone to scarce data problems. To avoid such issues, we have decoupled the analysis of speaker and verbal content information, and used a light-weight component to tackle each: an SVD component for global spectral matching, and a DTW component for local temporal alignment. Because these two components complement each other well, their integration leads to a satisfactory performance for the (language-independent) task considered. An equal error rate figure of approximately 4% has been obtained in experiments including deliberate imposture attempts. This approach was commercially released in October 1999 as part of the "VoicePrint Password" feature of MacOS 9.

7. REFERENCES

[1] G. Doddington, "Speaker Recognition—Identifying People by their Voices," *Proc. IEEE*, Vol. 73, November 1985.

[2] J.P. Campbell Jr., "Speaker Recognition: A Tutorial," *Proc. IEEE*, Vol. 85, No. 9, pp. 1437–1462, September 1997.

[3] A. Higgins, L. Bahler, and J. Porter, *Digital Signal Processing*, Vol. 1, pp. 89–106, 1991.

[4] T. Matsui and S. Furui, "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition," in *Proc.* 1994 *ICASSP*, Adelaide, Australia, pp. 125–128, April 1994.

[5] S. Parthasaraty and A.E. Rosenberg, "General Phrase Speaker Verification Using Sub–Word Background Models and Likelihood–Ratio Scoring," in *Proc. ICSLP*, Philadelphia, PA, October 1996.

[6] Q. Li, B.-H. Juang, Q. Zhou, and C.-H. Lee, "Automatic Verbal Information Verification for User Authentification," *IEEE Trans. SAP*, Vol. 8, No. 5, pp. 585–596, September 2000.

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

[8] Y. Ariki and K. Doi, "Speaker Recognition Based on Subspace Methods," in *Proc. ICSLP*, Yokohama, Japan, pp. 1859–1862, September 1994.

[9] K. Assaleh *et al.*, "Text–Dependent Speaker Verification Using Data Fusion and Channel Detection," in *Proc. SPIE*, Vol. 2277, San Diego, CA, pp. 72–82, July 1994.

[10] J. Lindberg and M. Bloomberg, "Vulnerability in Speaker Verification—A Study of Technical Impostor Techniques," in *Proc. EuroSpeech*, Budapest, Hungary, pp. 1211–1214, September 1999.