Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 87 (2007) 500–508
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

A signal subspace approach for speech modelling and classification

Alan W.C. Tan*, M.V.C. Rao, B.S. Daya Sagar

Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia

Received 13 October 2005; received in revised form 5 April 2006; accepted 15 June 2006
Available online 28 July 2006

Abstract

In this paper, a speech classifier inspired by the signal subspace approach is developed. A novel signal subspace speech model is initially obtained via a rank reducing subspace decomposition algorithm that is based on the SVD. Motivated by the assumption that the speech signal comprises of short term dynamics that are slowly changing, it follows that the signal subspace of the speech signal is likewise slowly changing. The proposed signal subspace model aims to characterize the subspace dynamics using a family of subspace trajectories. In particular, each subspace trajectory is a sequence of vectors that traces the dynamics of a rank-one subspace in time. An assembly of these trajectories, henceforth, specifies the progression of the embedded signal subspace. To construct the signal subspace classifier, prototype elements in the form of the signal subspace models are determined for every signal class. A minimum-distance rule with a distance measure that resembles an energy difference function is subsequently applied in the actual classification task. Simulation of the proposed signal subspace classifier in an isolated digit speech recognition problem reveals promising results.

Keywords: Speech modelling; Speech recognition; Subspace methods; Classification

1. Introduction

Subspace methods in speech signal analysis commonly relate to the application of the singular value decomposition (SVD) or its variants to reveal the principal features of the underlying signal. The general premise that information in speech signals is almost completely contained in a rank deficient subspace of the signal matrix enables the SVD to function as an analysis tool to abstract the desired signal subspace. As the measured signal is usually corrupted with additive noise, retaining only the signal content of the rank deficient signal subspace lends it a certain noise filtering quality. These ideas have been extensively researched for speech enhancement techniques (e.g., [1–3]). Extension of the signal subspace approach to coloured noise removal (e.g., [4,5]) and the evaluation of the subspace-based speech enhancement for robust speech recognition (e.g., [6,7]) have also been reported in the literature. In this paper, we take the signal subspace analysis in a slightly different direction for a classification problem.

Speech signals are strictly nonstationary. The basic assumption that speech signals consist of short term dynamics that are slowly changing is often

*Corresponding author. Tel.: +606 2523345; fax: +606 2316552.
E-mail address: wctan@mmu.edu.my (A.W.C. Tan).

0165-1684/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2006.06.008

necessary in time domain speech modelling [8, Chapters 2–3; 9, Chapter 4]. In fact, popular techniques such as the short-time Fourier transform (STFT), linear prediction coding (LPC) [10–12], and cepstral methods [10,13,14], are developed on the basis that the spectral content is slowly changing across the entire speech signal. These methods generally function to specify the relevant acoustic events in the speech signal in terms of a compact and efficient set of speech parameters. Building on a similar assumption, we infer that the signal subspace of the speech signal is also slowly changing coinciding with these short term dynamics. The proposed signal subspace model is a means to specify these signal subspace dynamics. A collection of signal subspace models, each representing a distinct signal class, establishes the signal subspace classifier. In particular, the signal subspace classifier is constructed by means of assembling prototype elements (in the form of the proposed speech model) of every signal class and thereafter executing a minimum-distance rule for a given measure of dissimilarity. Preliminary results from the simulation of the proposed signal subspace classifier in an isolated digit speech recognition problem appear promising.

The rest of the paper is structured as follows: In Section 2, a brief review on subspace methods and SVD is presented. Section 3 lays the foundation of the proposed signal subspace model. Then, Section 4 develops the actual subspace classification strategy. Simulation results and discussions are recorded in Section 5. Finally, Section 6 provides the major conclusions.

2. Signal model and the SVD

The linear model for the *clean* speech signal assumes that each n-dimensional vector $\mathbf{s}$ of the signal can be represented as [2]

$$\mathbf{s} = H\mathbf{y} = \sum_{i=1}^{p} \mathbf{h}_i y_i, \quad p \leqslant n \tag{1}$$

where $H = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p] \in \mathbb{R}^{n\times p}$ is a model matrix whose columns are orthogonal basis vectors that span the *signal subspace* and $\mathbf{y} = [y_1, y_2, \ldots, y_p]^{\mathrm{T}}$ is a zero mean random coefficient vector. In general, it is always possible to adequately represent speech signals using only $p < n$ basis vectors, i.e., $\mathbf{s} \in \mathscr{R}(H) \subset \mathbb{R}^n$. Let $\mathbf{x}$ denote the noisy measurement vector such that

$$\mathbf{x} = \mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{n}$ denotes the vector of the noise process. The noise process is assumed white, zero mean, additive and uncorrelated with the clean signal.

In practice, one observes and constructs a matrix $X \in \mathbb{R}^{n\times m}$ of measurements [1,10]. Here, m is the number of measurement channels and n is the number of measurements over these channels. Typically, n is much larger than m. One method to construct this matrix from a realization consisting of K samples $x_1, x_2, \ldots, x_K$ is to arrange these samples into a $n \times m$ matrix with Toeplitz structure [3]

$$X = \begin{bmatrix} x_m & x_{m-1} & \cdots & x_1 \\ x_{m+1} & x_m & \cdots & x_2 \\ \vdots & \vdots & & \vdots \\ x_K & x_{K-1} & \cdots & x_{K-m+1} \end{bmatrix}, \tag{3}$$

where the matrix dimension is constrained by $K = n + m - 1$. Equivalently, a Hankel matrix could have been used as both matrix structures are interchangeable by a simple permutation of columns. Consequently, we have

$$X = S + N \quad \text{and} \quad S = HY, \tag{4}$$

where S and N are the signal and noise matrices, respectively, and $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m] \in \mathbb{R}^{p\times m}$ is the matrix of coefficients.

The thin SVD (or economy-size SVD) of the measurement matrix X is given by [15, p. 72]

$$X = U\Sigma V^{\mathrm{T}} = \sum_{k=1}^{m} \mathbf{u}_k \sigma_k \mathbf{v}_k^{\mathrm{T}}, \tag{5}$$

where the columns of $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m] \in \mathbb{R}^{n\times m}$ are mutually orthonormal, $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m] \in \mathbb{R}^{m\times m}$ is an unitary matrix and $\Sigma \in \mathbb{R}^{m\times m}$ has the form

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m). \tag{6}$$

The diagonal elements of Σ are the *singular values* of X and are ordered so that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_m$. The columns of U and V are, respectively, called the *left* and *right singular vectors*. The SVD of the measurement matrix is also useful to describe the eigendecomposition of the sample correlation matrix R_x defined as [3]

$$R_x \triangleq \frac{1}{n} X^{\mathrm{T}} X = \frac{1}{n} V\Sigma^2 V^{\mathrm{T}}. \tag{7}$$

Under the assumptions that noise is white and uncorrelated with the clean signal, the row space of the signal matrix S can be completely recovered from the measurement matrix X [1]. Specifically, the p right singular vectors corresponding to the largest singular values of X are precisely the right singular vectors of S. An approximation to the reduced p-rank signal correlation matrix R_s may, therefore, be obtained as

$$R_s \triangleq \frac{1}{n} S^T S \approx \frac{1}{n} \sum_{k=1}^{p} \sigma_k^2 \mathbf{v}_k \mathbf{v}_k^T. \tag{8}$$

By the SVD, a consistent estimate of the row space of S, or equivalently, the eigenspace of R_x, can be obtained from X. This decomposition procedure is similar, in nature, to a Karhunen–Loeve transform (KLT) on the noisy measurement vector. For the rest of the paper, the aforementioned assumptions are tacitly implied.

3. Subspace modelling

Speech signals are essentially nonstationary and consist mainly of short term dynamics that are slowly changing [8, Chapter 2]. The boundaries of these dynamics generally mark the start and end of a phoneme, the smallest speech unit. The nonstationary dynamics present in speech signals almost invariably necessitates speech processing applications to work with frames [9, Chapter 4]. The general supposition that the speech signal can be reasonably assumed to be stationary over a frame interval is often an implicit requirement in these applications. By the same token, we have pursued a frame-based approach in the analysis; that is, a running rectangular window is used to acquire the analysis frames of the speech signal. In particular, the window is K length and advances every K_1 samples. The tth frame would, therefore, consist of the samples $K_1(t-1)+1, K_1(t-1)+2, \ldots, K_1(t-1)+K$. There are, in total, $T = \lceil (L-K)/K_1 + 1 \rceil$ frames where L is the number of samples in the speech signal and the operator $\lceil x \rceil$ returns the smallest integer greater than x. The samples contained in each frame are, thereafter, organized into the Toeplitz structure resembling Eq. (3). Following this procedure, we thus obtain the set of measurement matrices $\{X(t) : t \in \mathbb{N}_T\} = \mathscr{T}(\mathbf{x})$ given the samples of signal $\mathbf{x}$. Here, $\mathscr{T}$ denotes the frame operator and $\mathbb{N}_i$ is the subset of natural numbers $\{1, 2, \ldots, i\}$.

The underlying assumption that the short term dynamics are slowly changing makes it possible to specify the speech signal as a composite of signal subspace dynamics which we will denote as the *family of subspace trajectories* Ψ. Since the row space of the signal matrix is retrievable from the measurements, the signal subspace implied hereafter relates to the row space in particular. Formally, the *subspace trajectory* $\psi(t)$ is defined as a vector-valued function that is nonzero (or *active*) in some interval $t_1 \leqslant t \leqslant t_2$. Each trajectory describes a particular sequence of the right singular vectors; collectively, the family of subspace trajectories characterizes the entire signal subspace. Two right singular vectors $\mathbf{v}_t$ and $\mathbf{v}_{t+1}$ of successive frames belong to the same trajectory if $|\cos^{-1}(\mathbf{v}_t^T \mathbf{v}_{t+1})| \leqslant \theta_{th}$, where $\theta_{th} < \pi/4$ is the *transition bound*. Incidentally, the upper bound on θ_{th} is required to ensure no more than one vector $\mathbf{v}_{t+1}$ (from the set of orthonormal right singular vectors) lies within θ_{th} of vector $\mathbf{v}_t$. The subspace trajectory decomposition algorithm is described next.

Algorithm 1. Subspace trajectory decomposition algorithm.

(1) Given signal $\mathbf{x}$, construct the set of measurement matrices $\{X(t) : t \in \mathbb{N}_T\} = \mathscr{T}(\mathbf{x})$.
(2) a. Obtain the set of singular values $\{\sigma_k : k \in \mathbb{N}_m\}$ and the set of right singular vectors $\{\mathbf{v}_k : k \in \mathbb{N}_m\}$ from the SVD of $X(1)$.
    b. For $k = 1$ to m, set $\psi_k(1) \leftarrow \sigma_k \mathbf{v}_k$.
    c. Set $\mathscr{I} \leftarrow \mathbb{N}_m$, $\mathscr{I}' \leftarrow \emptyset$ and $M \leftarrow m$.
(3) For $t = 2$ to T, do:
    a. Obtain the set of singular values $\{\sigma_k : k \in \mathbb{N}_m\}$ and the set of right singular vectors $\{\mathbf{v}_k : k \in \mathbb{N}_m\}$ from the SVD of $X(t)$.
    b. For $k = 1$ to m, do:
        i. For $\forall i \in \mathscr{I}$, determine $\theta_i \leftarrow |\cos^{-1}(\frac{\mathbf{v}_k^T \psi_i(t-1)}{|\psi_i(t-1)|})|$.
        ii. Find $\theta_0 \leftarrow \min_{i \in \mathscr{I}}\{\theta_i\}$ and $i_0 \leftarrow \arg\min_{i \in \mathscr{I}}\{\theta_i\}$.
        iii. If $\theta_0 \leqslant \theta_{th}$, then set $\psi_{i_0}(t) \leftarrow \sigma_k \mathbf{v}_k$ and $\mathscr{I}' \leftarrow \mathscr{I}' \cup i_0$. Otherwise set $\psi_{M+1}(t) \leftarrow \sigma_k \mathbf{v}_k$, $\mathscr{I}' \leftarrow \mathscr{I}' \cup (M+1)$ and $M \leftarrow M+1$.
    c. Update $\mathscr{I} \leftarrow \mathscr{I}'$ and $\mathscr{I}' \leftarrow \emptyset$.
(4) Construct the family of subspace trajectories $\Psi \leftarrow \{\psi_j : j \in \mathbb{N}_M\}$.

The family of subspace trajectories Ψ obtained in Algorithm 1 consists of trajectories in both the

signal and noise subspace. It becomes necessary, therefore, to devise a selection strategy that discards all but the signal related trajectories of Ψ. In this respect, we define the energy E of trajectory $\psi \in \Psi$ as the sum of squared norms, i.e.,

$$E(\psi) = \sum_{t=1}^{T} \|\psi(t)\|^2. \tag{9}$$

Following the construction of ψ in Algorithm 1, the norms of ψ are also the singular values corresponding to the right singular vectors of the measurement matrices contained in ψ. It turns out, therefore, that the energy function accumulates the eigenvalues corresponding to the eigenvectors of the sample correlation matrices in ψ.

For a given *energy ratio* $E_{\text{th}} < 1$, the smallest subset $\Psi' \subset \Psi$ satisfying the constraint

$$\frac{\sum_{\psi \in \Psi'} E(\psi)}{\sum_{\psi \in \Psi} E(\psi)} > E_{\text{th}} \tag{10}$$

is denoted the *minimal set of subspace trajectories* $\breve{\Psi}$. To construct $\breve{\Psi}$, the trajectories in Ψ are ordered according to their energy and then picked for $\breve{\Psi}$, in a descending order of the energies, until the energy constraint is met.

Algorithm 2. Subspace trajectory selection algorithm.

(1) Given the family of subspace trajectories $\Psi = \{\psi_j : j \in \mathbb{N}_M\}$.
(2) For $j = 1$ to M, compute the energy map $E_j \leftarrow \sum_{t=1}^{T} \|\psi_j(t)\|^2$.
(3) Pick the minimal set of indices $\mathscr{J} \subset \mathbb{N}_M$ such that $\sum_{k \in \mathscr{J}} E_k > E_{\text{th}} \sum_{l \in \mathbb{N}_M} E_l$.
(4) Construct the minimal set of subspace trajectories $\breve{\Psi} \leftarrow \{\psi_j : j \in \mathscr{J}\}$.

Fig. 1 shows the dominant subspace trajectory of the utterance "Eight" and the labelling conventions used. Figs. 2 and 3 display the subspace trajectories, ranked in the order of decreasing energies, corresponding to the utterances "Eight" and "Five", respectively.

4. Subspace classification

4.1. The signal classification problem

Let $\mathscr{X}$ denote a signal space that contains all signals or time series and $\mathscr{C} = \{1, 2, \ldots, C\}$ be the set of class labels of C known classes. The central task of signal classification is the assignment of signals to

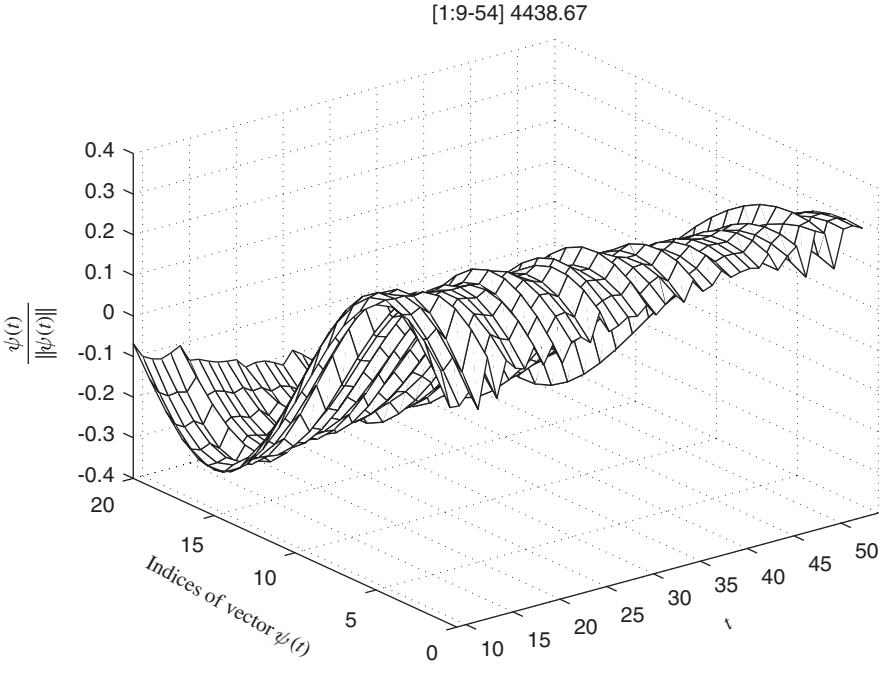

Fig. 1. The dominant subspace trajectory of the utterance "Eight", with $K = 160$, $K_1 = 40$, $m = 20$, $\theta_{\text{th}} = 25°$, $E_{\text{th}} = 0.9$. The plot is titled according to the convention "[i: t_1–t_2] E_i" where i denotes the ranking order.

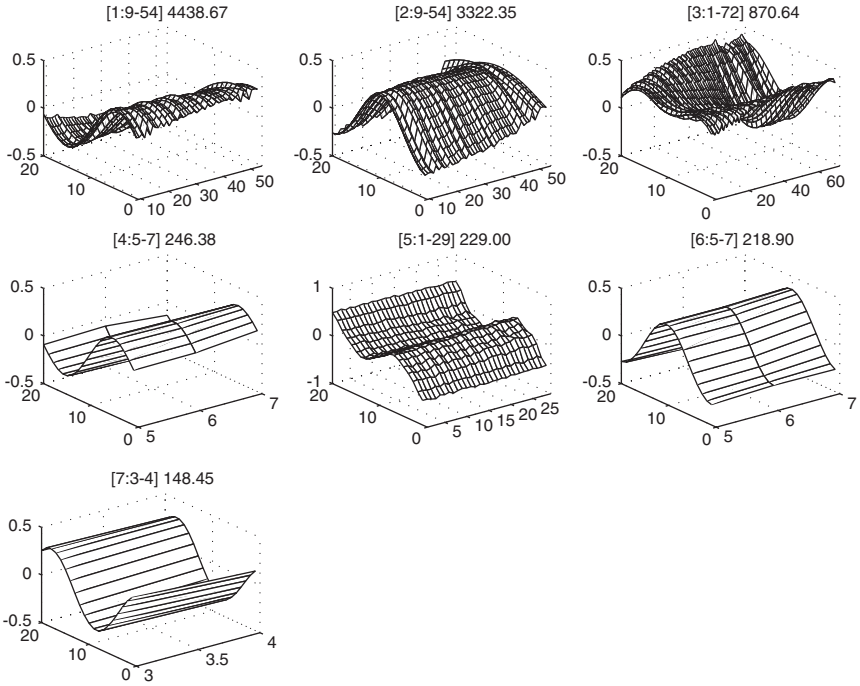

Fig. 2. Subspace trajectories of the utterance "Eight", with $K = 160$, $K_1 = 40$, $m = 20$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.9$.

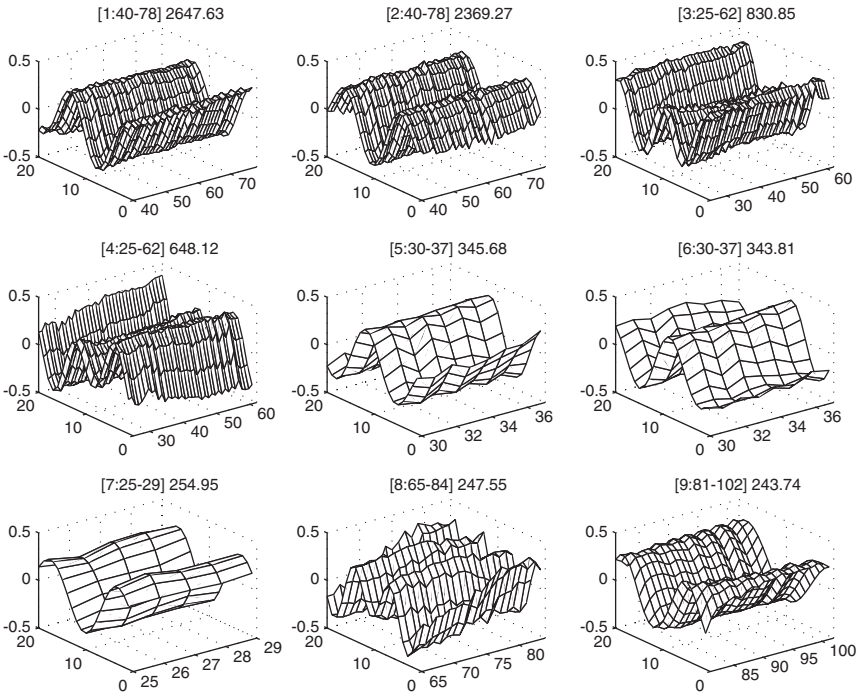

Fig. 3. Subspace trajectories of the utterance "Five", with $K = 160$, $K_1 = 40$, $m = 20$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.85$.

classes with strictly defined characteristics, or equivalently, the mapping $\mathscr{X} \to \mathscr{C}$. The signal space, however, is usually overly redundant and as a preliminary to the actual classification task, a *feature extractor* $f : \mathscr{X} \to \mathscr{F}$ is typically employed to reduce the dimensionality of the signal space. The

feature space $\mathscr{F}$ should essentially contain only relevant features of the signal under consideration. The final classification step is the map $g : \mathscr{F} \to \mathscr{C}$.

The final classification step may be considered as the map of distinct partitions of $\mathscr{F}$ onto some particular class label, i.e., $\mathscr{F}$ is partitioned into distinct regions $\{\mathscr{F}_c, c \in \mathscr{C}\}$ such that $g : \mathscr{F}_c \to c$. We may then associate a *prototype* as the representative element for the region $\mathscr{F}_c \subset \mathscr{F}$ such that an input $\mathbf{x} \in \mathscr{X}$ gets the class label of the prototype to which $\mathbf{x}$ is similar. The notion of similarity is generally based on a certain *distance measure*. The smaller the distance, the higher the degree of similarity between $\mathbf{x}$ and the prototype.

4.2. Subspace classification

To place the signal subspace analysis within the scope of a signal classification problem, several similarities have to be drawn. The speech model Ψ obtained in Algorithm 1 is a mapping of the signal $\mathbf{x}$ onto the row space of its measurement matrices. This transformation is, in essence, rank reducing and not invertible since the information contained in the column space of the measurement matrices is lost in the decomposition process. Relevant features of $\mathbf{x}$ are subsequently extracted from Ψ by Algorithm 2 and, with that, the noise-like features of $\mathbf{x}$ are removed. The resultant speech model $\breve{\Psi}$ is then taken to be the prototype element of its class. The two algorithms, therefore, collectively execute the role of the feature extractor, i.e., $\breve{\Psi} = f(\mathbf{x})$.

For every class $c \in \mathscr{C}$, let $\breve{\Psi}_c = \{\psi_{cj} : j \in \mathscr{J}_c\}$ be the prototype element of the class. Given an unknown signal $\mathbf{x}_0$, we desire to classify it into one of the classes in $\mathscr{C}$. To that end, we define a distance measure of the form

$$d(\psi_{cj}, X) \triangleq |\|\psi_{cj}\| - \|X\psi_{cj}\|/\|\psi_{cj}\||, \tag{11}$$

where $X \in \mathscr{T}(\mathbf{x}_0)$.

Lemma 1. *If $\mathbf{x}_0$ is the signal that generates the class prototype $\breve{\Psi}_c$, i.e., $\breve{\Psi}_c = f(\mathbf{x}_0)$, then $d(\psi_{cj}, X) = 0$ for every $\psi_{cj} \in \breve{\Psi}_c$ and $X \in \mathscr{T}(\mathbf{x}_0)$ of the same frame.*

Proof. Let the SVD of the measurement matrix be $X = \sum_{k=1}^{m} \mathbf{u}_k \sigma_k \mathbf{v}_k^{\mathrm{T}}$ and $\psi_{cj} = \sigma_{cj}\mathbf{v}_{cj}$. Then

$$X\psi_{cj}/\|\psi_{cj}\| = \sum_{k=1}^{m} \mathbf{u}_k \sigma_k \cos\,\phi_{cjk}, \tag{12}$$

where $\cos\,\phi_{cjk} = \mathbf{v}_k^{\mathrm{T}}\mathbf{v}_{cj}$. The norm of Eq. (12) gives

$$\|X\psi_{cj}\|/\|\psi_{cj}\| = \sqrt{\sum_{k=1}^{m} \sigma_k^2 \cos^2 \phi_{cjk}} \tag{13}$$

since the left singular vectors of X are mutually orthogonal. If $\mathbf{x}_0$ generates $\breve{\Psi}_c$ then $\mathbf{v}_{cj}$ must be exactly one of the right singular vectors of X and σ_{cj} is the corresponding singular value, i.e., there exists an l such that $\mathbf{v}_l = \mathbf{v}_{cj}$ and $\sigma_l = \sigma_{cj}$. Consequently, $\cos\,\phi_{cjl} = 1$ and $\cos\,\phi_{cjk} = 0$ for $k \neq l$ and hence $\|X\psi_{cj}\|/\|\psi_{cj}\| = \sigma_{cj} = \|\psi_{cj}\|$. □

In a sense, the norm in Eq. (13) may be interpreted as a measure of the content of X in ψ_{cj}. The deviation of this measure from the actual signal content in ψ_{cj} gives the desired distance measure of Eq. (11).

To classify the unknown signal $\mathbf{x}_0$, the weighted average $\mathscr{A}_w$ of the distance measure is evaluated across all trajectories in $\breve{\Psi}_c$, i.e.,

$$\mathscr{A}_w(\breve{\Psi}_c, \mathbf{x}_0) = \frac{\sum_{j\in\mathscr{J}_c}\sum_{t=1}^{T} w_{cj}(t)\, d(\psi_{cj}(t), X(t))}{\sum_{j\in\mathscr{J}_c}\sum_{t=1}^{T} w_{cj}(t)} \tag{14}$$

One particular choice for the weighting coefficients w is the norms of the trajectories, i.e., $w_{cj}(t) = \|\psi_{cj}(t)\|$. This choice of w tends to bias towards trajectories with higher signal content. A minimum-distance rule is subsequently applied to pick the class label of the prototype most similar to $\mathbf{x}_0$.

Algorithm 3. Subspace classification algorithm.

(1) Given unknown signal $\mathbf{x}_0$ and prototypes $\{\breve{\Psi}_c : c \in \mathscr{C}\}$.
(2) For $c = 1$ to C, set $v_c \leftarrow \mathscr{A}_w(\breve{\Psi}_c, \mathbf{x}_0)$ according to Eq. (14).
(3) Classify $\mathbf{x}_0$ according to the minimum-distance rule, i.e., $\arg\min_{c\in\mathscr{C}}\{v_c\}$.

5. Results and discussions

This section examines the performance of the proposed signal subspace classifier as a speech recognizer in two test cases.

5.1. Isolated digit speech recognition

Speech recordings, at the sampling frequency 10 kHz and with a signal-to-noise ratio (SNR) of approximately 24 dB, are collected from a pool of three male speakers. For every digit between 1–9

and two different utterances of the digit 0, i.e., "Zero" and "Oh", ten recordings are obtained, thereby yielding a total of $11 \times 10 = 110$ recordings per speaker. Next, a recording of each digit is selected randomly to build the set of class prototypes according to Algorithms 1 and 2 while the rest (nine recordings) are used as the testing data. The signal subspace classifier obtained is then evaluated according to Algorithm 3 for the testing data with the average recognition rate (of the three speakers) as the yardstick for performance. Following [3], we have chosen $K = 160$ and $K_1 = 40$ in our experiment. Main results of the simulation are shown in Table 1.

Table 1
Recognition rate, in percentage (%), and the mean number of trajectories per subspace prototype for various configurations of the signal subspace classifier

Classifier	Recognition rate (%)	Number of trajectories
$m = 20$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.9$	85.2	28.8
$m = 20$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.75$	75.1	8.6
$m = 20$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.95$	85.2	54.4
$m = 20$, $\theta_{\text{th}} = 15^\circ$, $E_{\text{th}} = 0.9$	85.2	67.5
$m = 20$, $\theta_{\text{th}} = 30^\circ$, $E_{\text{th}} = 0.9$	83.8	20.6
$m = 16$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.9$	81.2	18.7
$m = 28$, $\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.9$	88.9	75.7
$m = 28$, $\theta_{\text{th}} = 35^\circ$, $E_{\text{th}} = 0.925$	89.6	47.0

The performance of the signal subspace classifier generally improves as the energy ratio E_{th} increases (Fig. 4a). The drawback in increasing E_{th}, however, is that more trajectories are produced in the decomposition procedure (Fig. 4b) and consequently the computation cost is raised. One way of reducing the trajectory count is by increasing the transition bound θ_{th}. It is observed in our experiments that the performance of the classifier is rather insensitive to moderate values of θ_{th}, i.e., $15^\circ \leqslant \theta_{\text{th}} \leqslant 30^\circ$. In contrast, the classifier is highly sensitive to the row space dimension m. Increasing m enhances the performance (Fig. 5a) but at the same time it also elevates the computation cost (Fig. 5b). In our experiments, we found that setting $m = 20$, $\theta_{\text{th}} = 25^\circ$ and $E_{\text{th}} = 0.9$ produces a classifier (dubbed Sub1) that is low on computation cost and moderate in performance. If we allow some compromise on the computation cost, then using

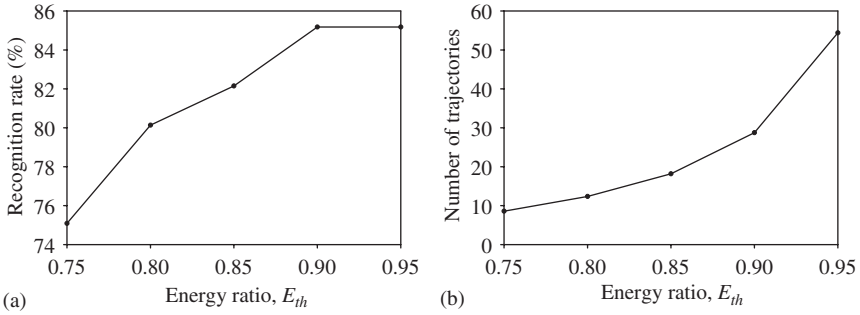

Fig. 4. (a) Performance of the signal subspace classifier, and (b) the mean number of trajectories per subspace prototype, as the energy ratio E_{th} varies ($m = 20$, $\theta_{\text{th}} = 25^\circ$).

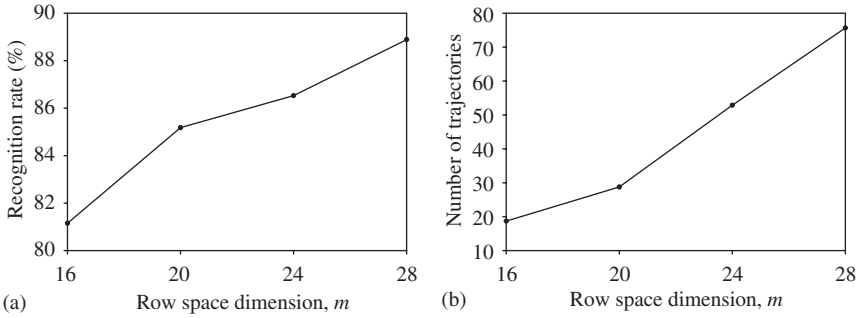

Fig. 5. (a) Performance of the signal subspace classifier, and (b) the mean number of trajectories per subspace prototype, as the row space dimension m varies ($\theta_{\text{th}} = 25^\circ$, $E_{\text{th}} = 0.9$).

$m = 28$, $\theta_{\text{th}} = 35^\circ$ and $E_{\text{th}} = 0.925$ yields a much improved classifier (dubbed Sub2).

For purposes of comparison, the same experimental setup is repeated for two commonly used speech recognizers, i.e., the LP-derived cepstral coefficients (LPCC) recognizer and the Mel-frequency-derived cepstral coefficients (MFCC) recognizer, both implementing dynamic time warping. Two variants of the LPCC, each using a different distortion measure, are tested: LPCC1 for the Euclidean distance and LPCC2 for the cepstral projection measure [16]. For all three recognizers tested, a Hamming window ($K = 240, K_1 = 80$) is applied to the data and 12 cepstral coefficients, liftered with $w_{\text{lift}}(k) = 1 + 6\sin(\pi k/12)$, are retained as the cepstral vector [8, Chapter 4]. The two LPCC recognizers achieve 89.2% (LPCC1) and 92.9% (LPCC2) recognition rate while the MFCC recognizer achieves a moderate 85.2%.

5.2. White noise robustness

The second part of our experiment investigates the white noise robustness of the proposed classifier. Artificial stationary white noise is added to the testing data and the classifier's performance at various levels of SNR is recorded [16]. Here, the SNR is defined on the entire utterance with the noise power assumed constant throughout the utterance. It is apparent from the results obtained (Fig. 6) that the proposed classifiers (Sub1 and Sub2) are remarkably robust to additive white noise. This is in contrast to the performance of the LPCC and MFCC recognizers which deteriorate in noisy environments. The white noise robustness of the proposed classifier owes much to the noise filtering quality inherent in subspace-based methods.

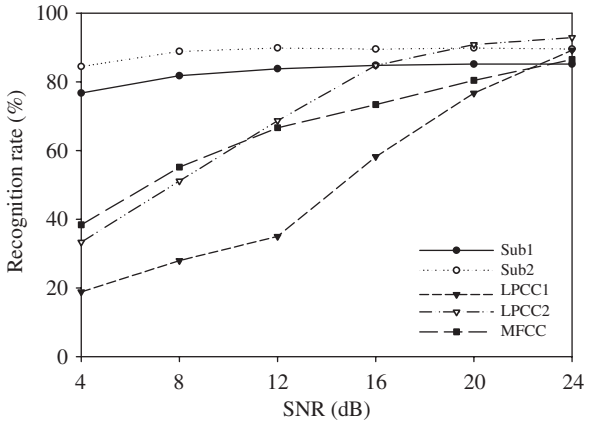

Fig. 6. Performance of various speech recognizers as SNR varies.

5.3. Implementation issues

Computational load, which is largely due to the SVD computation, remains an important issue in subspace processing. The direct application of frame-to-frame SVD computation is, however, rarely used in practice. Instead, efficient techniques which looks into approximating the SVD by alternative decomposition techniques that are cheaper computation-wise are more frequently used. Such approach generally features a mechanism that updates the estimates at each successive frame. Some examples are [4,19].

We have not mentioned explicitly how the parameters m, θ_{th} and E_{th} should be chosen. Raising either m or E_{th} generally improves the recognition rate (Figs. 4a and 5a) but at the expense of increased computation cost (Figs. 4b and 5b). Choosing E_{th} too large, additionally, may cause the subspace model to overfit the training speech data and thereby sacrificing robustness. In our work, we obtain a reasonable compromise between recognition accuracy and processing time using Sub1 and Sub2. For θ_{th}, it is difficult to ascertain, at this preliminary stage, how it actually affects the recognition task. The choice of a larger θ_{th} tends to generate fewer trajectories because there is a larger tolerance for variation in the right subspace vectors while a smaller θ_{th} generates more trajectories for the opposite reason. Without compromising too strongly on either the recognition rate or processing time, the suggested value for θ_{th} in either Sub1 or Sub2 achieves reasonably good results.

5.4. Future work

It is interesting to note that most misclassification involving the following digits occurs for the following cases:[1] "Five" as "Nine" (87.9%), "Nine" as "Five" (88.5%) and "Zero" as "Four" (82.8%). Predictably, these observations reveal that a great percentage of misclassification occurs in the sets of digits that are roughly phonetically alike. This is due, to a large extent, to the absence of a measure of discrimination in the selection strategy of Algorithm 2. In particular, the "best representation" approach

[1]The figure in parenthesis denote the percentage of the particular misclassification over all misclassification of the class.

taken by Algorithm 2 lacks a proper measure that evaluates the discriminative quality of the elements in the prototype set. Citing an parallel in the best-basis algorithm of [17], the original best-basis algorithm, which is a best representation realization of a given time series, has been adapted for classification problems by integrating discriminative qualities in the modelling strategy (e.g., [18]). In view of that, we surmise that the proposed signal subspace classifier could further be improved by adopting the same measure.

Extension of the proposed classifier to coloured noise removal is also possible following the approach taken by [3,4]. In particular, if the addictive noise is coloured, a prewhitening transformation is applied to the measurement matrix. The quotient SVD (QSVD) then replaces the SVD as a means to extract the signal subspace components of the speech signal. Another possible research direction is the application of the signal subspace features extracted by the methods presented here in training hidden Markov model (HMM) based speech recognizers. Our future work will address these important issues.

Although the signal subspace classifier is only comparable in performance to the industry standard speech recognizers, like the HTK Toolkit [20] which is based on the HMM, the application of the proposed subspace classifier in an isolated digit speech recognition problem is achieved with promising results.

6. Conclusion

A speech classifier inspired by the signal subspace approach is proposed in this paper. Class prototypes are constructed following a subspace decomposition algorithm that maps a signal or time series into a family of distinct subspace trajectories and, thereafter, retains a minimal set of trajectories that satisfies an energy constraint. The actual classification task is accomplished by a minimum-distance rule that picks the class label of the prototype that minimizes a particular energy difference function. Finally, the signal subspace classifier is successfully simulated in an isolated digit speech recognition problem with promising results.

Acknowledgements

The authors would like to thank the anonymous reviewers for the critical reading and insightful comments that have greatly helped to improve this paper's overall presentation.

References

[1] B. De Moor, The singular value decomposition and long and short spaces of noisy matrices, IEEE Trans. Signal Process. 41 (9) (1993) 2826–2838.
[2] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement, IEEE Trans. Speech Signal Process. 3 (4) (1995) 251–266.
[3] P.S.K. Hansen, Signal subspace methods for speech enhancement, Ph.D. Thesis, Technical University of Denmark, 1997.
[4] S.H. Jensen, P.C. Hansen, S.D. Hansen, J.A. Sørensen, Reduction of broad-band noise in speech by truncated QSVD, IEEE Trans. Speech Audio Process. 3 (6) (1995) 439–448.
[5] H. Lev-Ari, Y. Ephraim, Extension of the signal subspace speech enhancement approach to colored noise, IEEE Signal Process. Lett. 10 (4) (2003) 104–106.
[6] J. Huang, Y. Zhao, Energy-constrained signal subspace method for speech enhancement and recognition, IEEE Signal Process. Lett. 4 (10) (1997) 283–285.
[7] K. Hermus, P. Wambacq, Assessment of signal subspace based speech enhancement for noise robust speech recognition, Proc. IEEE ICASSP 1 (2004) 945–948.
[8] L.R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[9] L.R. Rabiner, R.W. Schaefer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.
[10] J.W. Picone, Signal modeling techniques in speech recognition, Proc. IEEE 81 (9) (1993) 1215–1247.
[11] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (4) (1975) 561–580.
[12] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech, Signal Process. 23 (1) (1975) 67–72.
[13] A.V. Oppenheim, R.W. Schafer, Homomorphic analysis of speech, IEEE Trans. Audio Electroacoust. 16 (2) (1968) 221–226.
[14] Y. Tokhura, A weighted cepstral distance measure for speech recognition, IEEE Trans. Acoustics Speech Signal Process. 35 (10) (1987) 1414–1422.
[15] G.H. Golub, C. Van Loan, Matrix Computations, second ed., John Hopkins University Press, Baltimore, MD, 1989.
[16] D. Mansour, B.H. Juang, A family of distortion measures based upon projection operation for robust speech recognition, IEEE Trans. Acoust. Speech Signal Process. 37 (11) (1989) 1659–1671.
[17] R.R. Coifman, M.V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Trans. Inform. Theory 38 (2) (1992) 713–718.
[18] N. Saito, R.R. Coifman, Local discriminant bases, Proc. SPIE Wavelet Applicat. Signal Image Process. II 2303 (1994) 2–14.
[19] E.C. Real, D.W. Tufts, J.W. Cooley, Two algorithms for fast approximate subspace tracking, IEEE Trans. Signal Process. 47 (7) (1999) 1936–1945.
[20] S.J. Young, P.C. Woodland, W.J. Byrne, HTK: Hidden Markov Model Toolkit V1.5, Cambridge University Engineering Department, Speech Group and Entropic Research Laboratories Inc., 1993.