Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Signal Processing ▌(▌▌▌▌) ▌▌▌–▌▌▌
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

A composite signal subspace speech classifier

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 24 November 2006; received in revised form 24 February 2007; accepted 19 April 2007

Abstract

Recently, a speech model inspired by signal subspace methods was proposed for a speech classifier. In using subspace information to characterize the speech signal, subspace trajectories in the form of the right singular vectors of the measurement matrices are obtained. Signal classification is thereafter accomplished by a minimum-distance rule with noteworthy results. This paper extends the foregoing approach by organizing the vector trajectories into matrices. The matrices so obtained are the reduced-rank approximation of the sample correlation matrices. A new dissimilarity measure in the Frobenius norm is correspondingly proposed for the matrix trajectories. Simulation results of the proposed composite signal subspace classifier in an isolated digit speech recognition problem reveal an improved performance over its predecessor. Additionally, the results also show the proposed classifier retaining the white noise robustness of the original design.

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

1. Introduction

The underlying assumption in signal subspace speech modelling lies in the premise that speech signals are strictly nonstationary and consist of short term dynamics that are slowly changing [1, Chapters 2 and 3; 2, Chapter 4]. In fact, popular speech processing techniques like the short-time Fourier transform (STFT), linear prediction coding (LPC) [3–5], and cepstral methods [3,6,7], are developed on the basis that the spectral content is slowly changing across the entire speech signal. These methods generally function by specifying the relevant acoustic events in the speech signal in terms of a compact and efficient set of speech parameters.

A signal subspace speech model is a characterization of the speech signal in terms of its subspace information. The signal subspace approach to speech processing was originally applied for speech enhancement techniques (such as [8–10]) and it has only been used recently as a speech classifier [11,12]. This paper extends the work in [11] for a slightly different class of signal subspace classifiers. In [11], subspace trajectories are characterized by the right singular vectors of the measurement matrices and the difference of the signal content in the trajectory set from its actual value is thereafter used as a token of the distance measure in the classification procedure. We propose here an alternative strategy, through organizing these singular vectors into matrices. It will be shown that the matrices obtained

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

0165-1684/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2007.04.009

in such a manner are the reduced-rank approximation of the sample correlation matrices. A new dissimilarity measure in the Frobenius norm is subsequently proposed for the matrix trajectories. The performance of the proposed composite signal subspace classifier is tested against its predecessor in an isolated digit speech recognition problem and results appear promising.

The rest of the paper is structured as follows: a brief review on signal subspace modelling is covered in Section 2. Section 3 develops the main ideas behind the proposed composite signal subspace classifier. Simulation results and discussions are recorded in Section 4 and Section 5 provides the major conclusions.

2. Subspace modelling

A running rectangular window is used to acquire the analysis frames of the speech signal. The window is K length and advances every K_1 samples. The tth frame, therefore, consists of the samples $K_1(t-1)+1, K_1(t-1)+2, \ldots, K_1(t-1)+K$ and 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 or equal to x. The samples contained in each frame, say $x_1, x_2, \ldots, x_K$, are then organized into a measurement matrix of the form [8,10]

$$X = \begin{bmatrix} x_m & x_{m-1} & \cdots & x_1 \\ x_{m+1} & x_m & \cdots & x_2 \\ \vdots & \vdots & & \vdots \\ x_K & x_{K-1} & \cdots & x_{K-m+1} \end{bmatrix} \in \mathbb{R}^{n \times m}, \tag{1}$$

where $n > m$ and the matrix dimension constrained by $K = n + m - 1$. Following this procedure, we obtain the set of measurement matrices $\{X(t): t \in \mathbb{N}_T\} = \mathscr{T}(\mathbf{x})$ of the 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 thin SVD (or economy-size SVD) of the measurement matrix X in (1) is defined as [13, p. 72]

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

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 a unitary matrix and $\Sigma \in \mathbb{R}^{m \times m}$ has the form

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

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 [10]

$$R_x = \frac{1}{n} X^{\mathrm{T}} X = \frac{1}{n} \sum_{k=1}^{m} \sigma_k^2 \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}}. \tag{2}$$

Formally, the *subspace trajectory* $\psi(t)$ is defined as a vector-valued function of the right singular vectors of successive measurement matrices [11]. It is nonzero (or *active*) in some frame interval $t_1 \leqslant t \leqslant t_2$ and two right singular vectors $\mathbf{v}_t$ and $\mathbf{v}_{t+1}$ of successive frames belong to the same trajectory if $|\cos^{-1}(\mathbf{v}_t^{\mathrm{T}} \mathbf{v}_{t+1})| \leqslant \theta_{th}$, where $\theta_{th} < \cos^{-1}(m^{-1/2})$ is the *transition bound*. Collectively, the *family of subspace trajectories* $\Psi = \{\psi_j(t): j \in \mathscr{J}, t \in \mathbb{N}_T\}$, where $\mathscr{J}$ represents the set of trajectory indices, characterizes the entire signal subspace.

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 $\mathscr{J} \leftarrow \mathbb{N}_m$, $\mathscr{J}' \leftarrow \emptyset$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 $\forall i \in \mathscr{J}$, determine $\theta_i \leftarrow \cos^{-1}\left|\frac{\mathbf{v}_k^{\mathrm{T}} \psi_i(t-1)}{\|\psi_i(t-1)\|}\right|$.
        (ii) Find $i_0 \leftarrow \operatorname{argmin}_{i \in \mathscr{J}}\{\theta_i\}$.
        (iii) If $\theta_{i_0} \leqslant \theta_{th}$, then set $\psi_{i_0}(t) \leftarrow \sigma_k \mathbf{v}_k$ and $\mathscr{J}' \leftarrow \mathscr{J}' \cup i_0$. Otherwise set $\psi_{M+1}(t) \leftarrow \sigma_k \mathbf{v}_k$, $\mathscr{J}' \leftarrow \mathscr{J}' \cup (M+1)$ and $M \leftarrow M+1$.
    (c) Update $\mathscr{J} \leftarrow \mathscr{J}'$ and $\mathscr{J}' \leftarrow \emptyset$.
(4) Construct the family of subspace trajectories $\Psi \leftarrow \{\psi_j: j \in \mathbb{N}_M\}$.
```

As a consequence to the rank degeneracy of the measurement matrices, it becomes necessary to pick a *minimal set of subspace trajectories* $\check{\Psi} \subset \Psi$ that retains only the dominant trajectories of Ψ [11]. For this purpose, an *energy ratio* E_{th} is predetermined and we have the following algorithm.

Algorithm 2. Subspace trajectory selection algorithm.

(1) Given the family of subspace trajectories $\Psi = \{\psi_j\colon\ j \in \mathbb{N}_M\}$.
(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 $\sum_{k\in\mathscr{J}} E_k > E_{th} \sum_{l\in\mathbb{N}_M} E_l$.
(4) Construct the minimal set of subspace trajectories $\check{\Psi} \leftarrow \{\psi_j\colon\ j \in \mathscr{J}\}$.

Figs. 1b and 2b show the extent of the individual trajectories in the minimal set of subspace trajectories for utterances "Eight" and "Five", respectively. The number in the parentheses indicates the energy of the corresponding trajectory.

3. Composite subspace trajectory

3.1. Definition

Another perspective of the subspace approach in signal modelling arises when the trajectory vectors are organized into matrices. Formally, we define the *composite subspace trajectory* $\acute{\Psi} = \{H(t)\colon t \in \mathbb{N}_T\}$ as a sequence of matrices of the form

$$H(t) = [\psi_1(t) \quad \psi_2(t) \quad \cdots \quad \psi_{p(t)}(t)],$$

where $\psi_1, \psi_2, \ldots, \psi_{p(t)}$ denote the active trajectories in frame t of a given set of subspace trajectories. We denote this transformation by the mapping $\mathscr{M}$.

From the definition of H, it is clear that the matrix HH^{T} is a scaled p-rank approximation of the sample correlation matrix R_x of the same frame, i.e.,

$$HH^{\mathrm{T}} = \sum_{i=1}^{p} \sigma_{l_i}^2 \mathbf{v}_{l_i} \mathbf{v}_{l_i}^{\mathrm{T}}, \tag{3}$$

where σ_{l_i} and $\mathbf{v}_{l_i}$ follow from the eigendecomposition of nR_x in (2) and $l_i \in \mathbb{N}_m$, $l_i \neq l_j$ unless $i = j$. This approximation is, however, not necessarily optimal in the 2-norm sense. The nearest 2-norm approximation of nR_x would have, for the indices of the sum in (3), $l_i = i$. In the case where the composite trajectory is obtained from the family of subspace trajectories generated by Algorithm 1, then $H \in \mathbb{M}^{m\times m}$ and $HH^{\mathrm{T}} = nR_x$. Otherwise, the composite trajectory is composed of distinct matrices of varying dimension and rank. Figs. 1c and 2c display the rank of these matrices for utterances "Eight" and "Five", respectively. There are, in total, 72 matrices characterizing the composite trajectory of the utterance "Eight" and 78 matrices for the utterance "Five".

As HH^{T} is, in essence, a rank p approximation of nR_x, it is conceivable to perceive a slightly different approach in obtaining $\acute{\Psi}$, i.e., by a direct minimal 2-norm reduced-rank approximation of the sample correlation matrices. To that end, an estimate of the rank must be determined for each of the sample correlation matrices. The model selection strategies employed in [14] and based on the information theoretic criteria of [15–17] are some of the common choices for this purpose. Figs. 1d and 2d show the rank of the matrices detected with the minimum description length (MDL) of [17]. From these plots, it is evident that the MDL models generate matrices of a significantly greater rank. As a result, these models are likely to overfit the signal subspace rendering the approach unsuitable for most classification tasks. Results from our experiments corroborate this observation.

3.2. Subspace classification

Let $\mathscr{C} = \{1, 2, \ldots, C\}$ denote the set of the C known classes. For every class $c \in \mathscr{C}$, let the class prototype $\acute{\Psi}_c = \{H_c(t)\colon t \in \mathbb{N}_T\}$ be the composite subspace trajectory of the signal class. Given an unknown signal $\mathbf{x}_0$, we desire to classify it into one of the classes in $\mathscr{C}$.

A straightforward dissimilarity measure that follows from the foregoing property of the composite subspace trajectory is

$$d(H_c, X) = \|X^{\mathrm{T}} X - H_c H_c^{\mathrm{T}}\|_{\mathrm{F}}^2$$

where $X \in \mathscr{T}(\mathbf{x}_0)$ and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. [1]As a consequence to (3), we have the following lemma.

Lemma 1. *If $\mathbf{x}_0$ is the signal that generates Ψ by Algorithm 1 and $\acute{\Psi}_c = \mathscr{M}(\Psi)$ is the corresponding composite subspace trajectory, then $d(H_c, X) = 0$ for every $H_c \in \acute{\Psi}_c$* and *$X \in \mathscr{T}(\mathbf{x}_0)$ of the same frame instant.*

[1] The Frobenius norm of a matrix $A \in \mathbb{R}^{n\times m}$ is defined as

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2},$$

where a_{ij} is the ijth element of the matrix [13, p. 55].

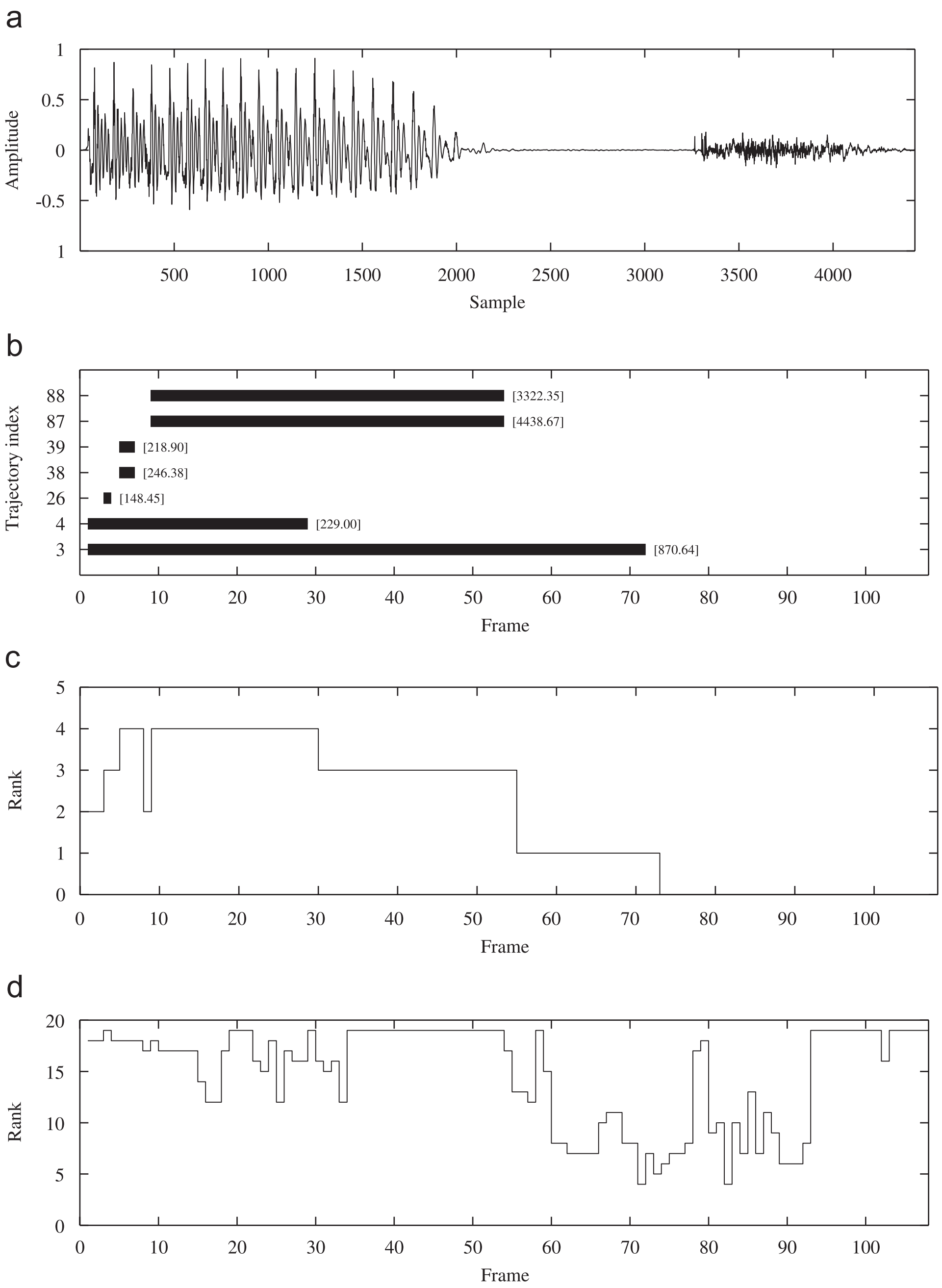

Fig. 1. (a) The utterance "Eight", (b) extent of the subspace trajectories, (c) rank of the composite subspace trajectory, and (d) rank of the MDL model ($K = 160$, $K_1 = 40$, $m = 20$, $\theta_{th} = 25^\circ$, $E_{th} = 0.9$).

The sum of dissimilarity measures over the entire utterance gives us a means to evaluate the likeness of the unknown signal with the prototype classes, i.e.,

$$\mathscr{A}(\acute{\Psi}_c, \mathbf{x}_0) = \sum_{t=1}^{T} d(H_c(t), X(t)) \tag{4}$$

Algorithm 3. Subspace classification algorithm.

(1) Given unknown signal $\mathbf{x}_0$ and prototypes $\{\acute{\Psi}_c : c \in \mathscr{C}\}$.
(2) For $c = 1$ to C, set $v_c \leftarrow \mathscr{A}(\acute{\Psi}_c, \mathbf{x}_0)$ according to (4).
(3) Classify $\mathbf{x}_0$ according to the minimum-distance rule, i.e., $\operatorname{argmin}_{c\in\mathscr{C}}\{v_c\}$.

Please cite this article as: A.W.C. Tan, et al., A composite signal subspace speech classifier, Signal Process. (2007), doi:10.1016/j.sigpro.2007.04.009

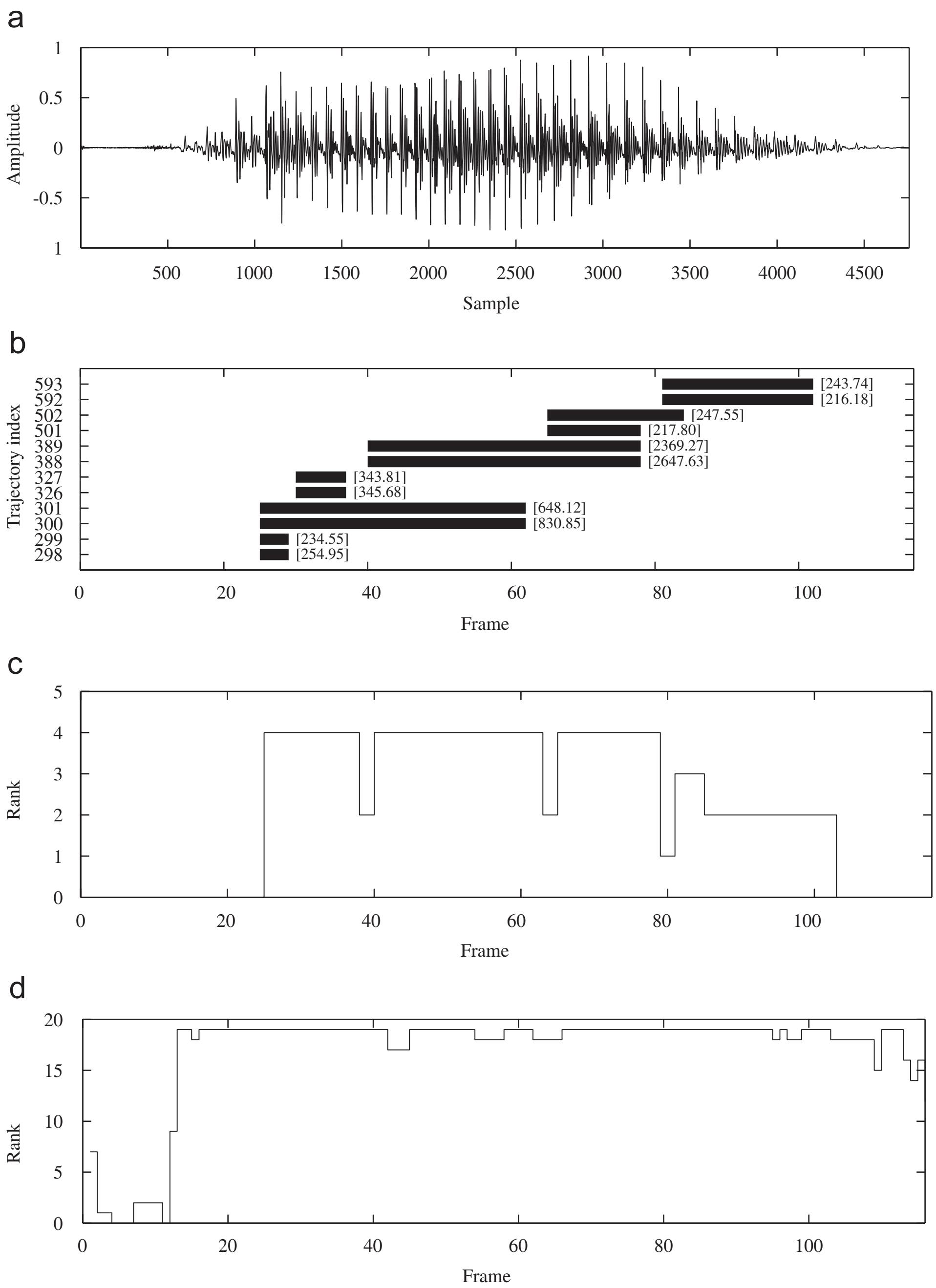

Fig. 2. (a) The utterance “Five”, (b) extent of the subspace trajectories, (c) rank of the composite subspace trajectory, and (d) rank of the MDL model ($K = 160$, $K_1 = 40$, $m = 20$, $\theta_{th} = 25°$, $E_{th} = 0.85$).

4. Results and discussions

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

4.1. Isolated digit speech recognition

Speech recordings, at a sampling frequency of 10 kHz, are collected from three male speakers. For every digit between 1 and 9 and two different utterances of the digit 0, i.e., “Zero” and “Oh”, 10 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 (Algorithms 1 and 2 and the mapping $\mathscr{M}$) while the other nine recordings are used as the testing data. The composite signal subspace classifier obtained is then evaluated on the testing data with the average recognition rate (of the three speakers) as the yardstick for performance.

In the simulation, we have used $K = 160$ and $K_1 = 40$ (following [10,11]), and chosen ($m = 20$, $\theta_{th} = 25^\circ$, $E_{th} = 0.9$) and ($m = 28$, $\theta_{th} = 35^\circ$, $E_{th} = 0.925$) as the parameter sets of two composite signal subspace classifiers, dubbed CSub1 and CSub2, respectively. For comparison purposes, we also presented simulation results on a few types of classifiers, namely, two subspace classifiers (Sub1 and Sub2) which implement the classification scheme in [11], two LP-derived cepstral coefficients (LPCC) recognizers and a Mel-frequency-derived cepstral coefficients (MFCC) recognizer. The two variants of the LPCC, each with a different distortion measure, are LPCC1 of the Euclidean distance and LPCC2 of the cepstral projection measure [18]. For the LPCC and MFCC recognizers, 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 [1, Chapter 4]. The main results of the simulation are displayed in Table 1.

It is apparent, from these results, that the proposed composite signal subspace classifiers (CSub1 and CSub2) are superior in performance to their predecessors (Sub1 and Sub2) and comparable to the LPCC and MFCC recognizers.

Table 1
Recognition rate, in percentage (%), of the various speech recognizers

Classifier	Recognition rate (%)
CSub1	92.3
CSub2	93.6
Sub1	85.2
Sub2	89.6
LPCC1	89.2
LPCC2	92.9
MFCC	85.2

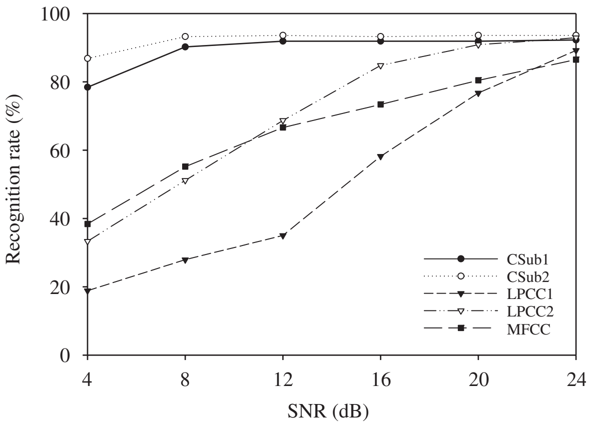

Fig. 3. Performance of various speech recognizers as SNR varies.

4.2. White noise robustness

The second part of our simulation investigates the white noise robustness of the proposed classifier. The signal-to-noise ratio (SNR) of the original recorded speech data is approximately 24 dB. Artificial stationary white noise is introduced into the testing data and the classifier’s performance at various levels of SNR is recorded [18]. As in [11], the proposed composite signal subspace classifier displays remarkable robustness to additive white noise (see Fig. 3). This is almost in complete contrast to the LPCC and MFCC recognizers which fare poorly in noisy environments. The white noise robustness of subspace classifiers is largely attributed to the noise filtering quality inherent in subspace-based methods.

5. Conclusion

This paper extends the work in [11] for a slightly different class of signal subspace classifiers. Through the organization of the right singular vectors of the measurement matrices into matrices, the composite subspace trajectory is obtained for the signal of interest. A new dissimilarity measure in the Frobenius norm that assesses the likeness of the sample correlation matrices with the composite subspace trajectory is correspondingly proposed. Results from simulation show that the proposed signal subspace classifier is superior to its predecessor in an isolated digit speech recognition problem, and at the same time, retains the white noise robustness of the original design.

Please cite this article as: A.W.C. Tan, et al., A composite signal subspace speech classifier, Signal Process. (2007), doi:10.1016/j.sigpro.2007.04.009

References

[1] L.R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[3] J.W. Picone, Signal modeling techniques in speech recognition, Proc. IEEE 81 (9) (September 1993) 1215–1247.

[4] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (4) (April 1975) 561–580.

[5] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 23 (1) (February 1975) 67–72.

[6] A.V. Oppenheim, R.W. Schafer, Homomorphic analysis of speech, IEEE Trans. Audio Electroacoust. 16 (2) (June 1968) 221–226.

[7] Y. Tokhura, A weighted cepstral distance measure for speech recognition, IEEE Trans. Acoustics Speech Signal Process. 35 (10) (October 1987) 1414–1422.

[8] B. De Moor, The singular value decomposition and long and short spaces of noisy matrices, IEEE Trans. Signal Process. 41 (9) (September 1993) 2826–2838.

[9] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process. 3 (4) (July 1995) 251–266.

[10] P.S.K. Hansen, Signal subspace methods for speech enhancement, Ph.D Thesis, Technical University of Denmark, 1997.

[11] A.W.C. Tan, M.V.C. Rao, B.S. Daya Sagar, A signal subspace approach for speech modelling and classification, Signal Process. 87 (3) (March 2007) 500–508.

[12] A.W.C. Tan, M.V.C. Rao, B.S. Daya Sagar, A discriminative signal subspace speech classifier, IEEE Signal Process. Lett. 14 (2) (February 2007) 133–136.

[13] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., The John Hopkins University Press, Baltimore, MD, 1996.

[14] M. Wax, T. Kailath, Detection of signals by information theoretic criteria, IEEE Trans. Acoust. Speech Signal Process. 33 (2) (April 1985) 387–392.

[15] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (6) (December 1974) 716–723.

[16] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[17] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.

[18] 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) (November 1989) 1659–1671.

Please cite this article as: A.W.C. Tan, et al., A composite signal subspace speech classifier, Signal Process. (2007), doi:10.1016/j.sigpro.2007.04.009