Expert Systems with Applications 38 (2011) 3690–3696

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Use of center of gravity with the common vector approach in isolated word recognition

M. Bilginer Gülmezoğlu[a,*], Rifat Edizkan[a], Semih Ergin[a], Atalay Barkana[b]

[a] Eskişehir Osmangazi University, Electrical and Electronics Engineering Department, Meşelik Campus, 26480 Eskişehir, Turkey
[b] Anadolu University, Electrical and Electronics Engineering Department, İki Eylül Campus, 26470 Eskişehir, Turkey

ARTICLE INFO

Keywords:
Center of gravity
Endpoint detection
Common vector approach
Speech recognition

ABSTRACT

In this paper, the subspace based classifier, common vector approach (CVA), with the center of gravity (COG) method is used for isolated word recognition. Since the CVA classifier is sensitive to shifts through the time axis, endpoint detection becomes extremely important for the recognition of isolated words. The COG method eliminates the need for endpoint detection. The effects of the COG method and a classical endpoint detection algorithm on the recognition rates of isolated words are investigated. The experimental results show that the COG method yields slightly higher recognition rates than the endpoint detection method in the TI-digit database when CVA is used.

1. Introduction

Isolated word recognition is the process of automatically extracting and then classifying the features conveyed by a speech wave using computers and electronic circuits. Automatic isolated word recognition methods have been investigated for many years aimed especially at phone dialing and commanding certain machinery including the computers.

The initial step in isolated word recognition is endpoint detection. Different algorithms for this purpose have been used for many years. Endpoint detection using zero-crossing and energy may not give satisfactory results especially when the word starts with a fricative sound (Rabiner & Sambur, 1975). A back propagation neural network has been used to recognize "non-speech" frames of the speech signal (Orság & Zbořil, 2003). Some of the endpoint detection methods mentioned in the literature are the classifier based methods. Orlandi, Santarelli, and Falavigna (2003) proposed a robust and computationally low-cost HMM-based start–endpoint detector for speech recognition. This classifier-based endpoint detector relies on general statistics rather than on local information, energy, and zero-crossings. In their work, a combination of energy and zero-crossing features, and Average Magnitude Difference Function (AMDF) and zero-crossings were all used as features. Most other endpoint detection methods are given for noisy environments (Huang & Yang, 2000; Li, Zheng, Tsai, & Zhou, 2002; Raj & Singh, 2003; Shen, Hung, & Lee, 1998; Shin, Lee, Lee, & Lee, 2000; Wu & Lin, 2000). Since our work contains only isolated word recognition in a noise-free environment, other endpoint detection methods in noisy environments will not be discussed here.

Failure in the endpoint detection degrades the recognition rates when subspace classifiers are used (Günal, Edizkan, & Barkana, 2003). Since these classifiers are sensitive to shifts through the time axis, the determination of the location of isolated word in any recording is important in subspace classifiers. For this purpose, the COG method is proposed for CVA subspace classifier in this paper. The COG method can almost precisely locate the center of gravity of the isolated word through the recording time interval when energy distribution along the word is considered. A specific number of samples to the right and left of the center of the word can be taken to approximately cover the word rather than calculating end-points.

Other researchers have investigated the COG idea for different purposes (Stylianou, 1998; Stylianou, 1999, 2001). Stylianou used the COG idea in concatenative speech synthesis for text-to-speech conversion. The technique for finding the center of gravity of speech signals is employed to synchronize speech frames and to obtain inter-frame coherence in speech coding and text-to-speech (TTS) synthesis applications (Stylianou, 1998). These are based on the concatenation of subword-sized units of recorded speech so that the phase mismatching between speech frames can be prevented (Stylianou, 1998). The center of gravity method is used in phase correction to change the position of the analysis window (Stylianou, 2001) and also as the first step in position reconstruction from a set of data (Landi, 2003). Feth, Fox, Jacewicz, and Iyer (2002) investigated the dynamic center of gravity effect observed in diphthongal vowel to consonant–vowel (CV) transition. Van Son and Pols (1996) studied the center of gravity of the spectrum as an aspect of vowels and consonants to characterize consonant reduction.

* Corresponding author. Tel.: +90 222 239 37 50x3261; fax: +90 222 229 05 35.
E-mail addresses: bgulmez@ogu.edu.tr (M.B. Gülmezoğlu), redizkan@ogu.edu.tr (R. Edizkan), sergin@ogu.edu.tr (S. Ergin), atalaybarkan@anadolu.edu.tr (A. Barkana).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.026

Each recording in the TI-digit database starts with a silence, continues with the utterance, and ends with another silence. In this study, COG is employed to find the center of the utterance from each recording in the TI-digit database. After taking a specific number of samples to the right and left of the center of the word, the root-melcep parameters are calculated for each frame of the extracted utterance. These parameters are used as the elements of the feature vectors. The recognition process is carried out using the common vector approach (CVA) (Çevikalp, Neamtu, Wilkes, & Barkana, 2005; Gülmezoğlu, Dzhafarov, Keskin, & Barkana, 1999, 2001).

COG with CVA is proposed in this paper since this methodology has several advantages over classical endpoint algorithm with CVA. The COG method approximately gives same regions in all words of each class since it is insensitive to time shift. Therefore the model parameters will be more representative for the class so that high recognition rates can be obtained from the isolated word recognition task. The proposed technique is applied only on word model based isolated word recognition in spite of the fact that endpoint detection in the speech recognition is already quite well handled implicitly during DTW and Viterbi recognition search processes.

2. Theory

In this section, the basic endpoint detection is first briefly reviewed and then COG in the time domain is introduced. The theoretical background of the CVA classifier will then be given for two different cases.

2.1. Endpoint detection

One approach for isolating speech from silence regions is the end-point detection (Rabiner & Sambur, 1975; Rabiner, 1978). The end-point detection algorithm used in this paper is based on two measurements for the isolated utterance: energy and zero-crossing rate (Rabiner & Sambur, 1975). In this algorithm, the speech signal is bandpass filtered to eliminate the low-frequency hum, the DC level, and the high frequency components. The energy and the zero-crossing rate are measured for every 12.5 ms of speech samples and their thresholds are calculated (Rabiner & Sambur, 1975) while no sound is present. The algorithm searches the beginning and end points of the utterance according to the energy thresholds. These estimated endpoints are then corrected with zero-crossing rates. The algorithm searches back a predetermined time and tries to locate the new estimate points based on zero-crossing thresholds. This algorithm is simple and fast because it uses integer arithmetic, but it may not give real endpoints every time.

2.2. COG formulas used in speech extraction

The COG of a signal can be defined as the COG of the energy distribution in time. Therefore, the COG can be used to locate the center of an isolated word almost precisely when the momentum through the word is considered. The whole word can be extracted by taking a certain number of samples to the right and left of the center of the word. This proposed COG method is an alternative to endpoint detection in the CVA-based isolated word recognition system.

Two formulas are applied to raw speech samples to find the COG (η) of the isolated word speech data:

$$\eta_1 = \frac{\sum_{n=0}^{\infty} n x^2(n)}{\sum_{n=0}^{\infty} x^2(n)} \tag{1}$$

$$\eta_2 = \frac{\sum_{n=0}^{\infty} n|x(n)|}{\sum_{n=0}^{\infty} |x(n)|} \tag{2}$$

where n denotes the sample number and $x(n)$ is the nth sample of the speech signal. Eqs. (1) and (2) are referred to as COG formula (1) and COG formula (2), respectively. These formulas can be found in any calculus book.

The COG method approximately gives the same frame corresponding to the same phoneme for all the words in one class. If the preceding and succeeding frames of this center frame are different for each of the classes, the classification task can be easily handled.

2.3. The common vector approach (CVA) used in the recognition process

CVA has been found to be more effective than other subspace classifiers according to our previous experience (Çevikalp et al., 2005; Gülmezoğlu et al., 1999, Gülmezoğlu, Dzhafarov, & Barkana, 2001, 2007). CVA has been studied for two different cases. One is the case when the number (m) of feature vectors is less than or equal to the dimension (p) of feature vectors ($m \leqslant p$). This case is called the insufficient data case. The second case occurs when the reverse happens ($m > p$) and it is called the sufficient data case. These two cases are explained below.

2.3.1. The insufficient data case

Let the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbf{R}^p$ be the feature vectors for a certain word-class c in the training set where $m \leqslant p$. Then each of these feature vectors, which are assumed to be linearly independent, can be written as

$$\mathbf{a}_i = \mathbf{a}_{i,dif} + \mathbf{a}_{com} + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, m \tag{3}$$

where the vector $\mathbf{a}_{i,dif}$ indicates inter- and intra-speaker differences as well as the acoustical environmental effects and the phase or temporal differences, and the vector $\mathbf{a}_{com}$ is the common vector of the word-class c, and ε_i represents the error vector (Gülmezoğlu et al., 2001). The common vector represents the common properties or invariant features of the word-class c. In Eq. (3), there are m vector equations with $(2m + 1)$ unknown vectors. Therefore there is an infinite number of solutions for $\mathbf{a}_{com}$, $\mathbf{a}_{i,dif}$ and ε_i ($i = 1, 2, \ldots, m$).

A unique solution for $\mathbf{a}_{com}$ can be obtained as (Gülmezoğlu et al., 2001)

$$\mathbf{a}_{com} = \mathbf{a}_i - \mathbf{a}_{i,dif}, \quad \forall i = 1, 2, \ldots, m \tag{4}$$

where

$$\mathbf{a}_{i,dif} = \langle \mathbf{a}_i, \mathbf{z}_1 \rangle \mathbf{z}_1 + \langle \mathbf{a}_i, \mathbf{z}_2 \rangle \mathbf{z}_2 + \cdots + \langle \mathbf{a}_i, \mathbf{z}_{m-1} \rangle \mathbf{z}_{m-1} \tag{5}$$

where $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m-1}\}$ constitutes an orthonormal basis vector set obtained from the difference vectors by using the Gram–Schmidt orthogonalization method.

The common vector does not depend on the choice of the orthonormal basis vector set of difference subspace $\mathbf{B}$ (Gülmezoğlu et al., 2001). Therefore, the common vector is unique for each class and all the error vectors ε_i would be zero.

In addition to this method, the common vector can also be obtained by using the covariance matrix. Let us define the covariance matrix of the feature vectors belonging to a word-class c as

$$\mathbf{\Phi} = \sum_{i=1}^{m} (\mathbf{a}_i - \mathbf{a}_{ave})(\mathbf{a}_i - \mathbf{a}_{ave})^T \tag{6}$$

The nonzero eigenvalues of the covariance matrix $\mathbf{\Phi}$ should correspond to the eigenvectors forming an orthonormal basis for the difference subspace $\mathbf{B}$ (Gülmezoğlu et al., 2001). The orthogonal complement $\mathbf{B}^{\perp}$ in this case is spanned by all the eigenvectors corresponding to the zero eigenvalues. This subspace is called the indifference subspace and has a dimension of $(p - m + 1)$. The direct

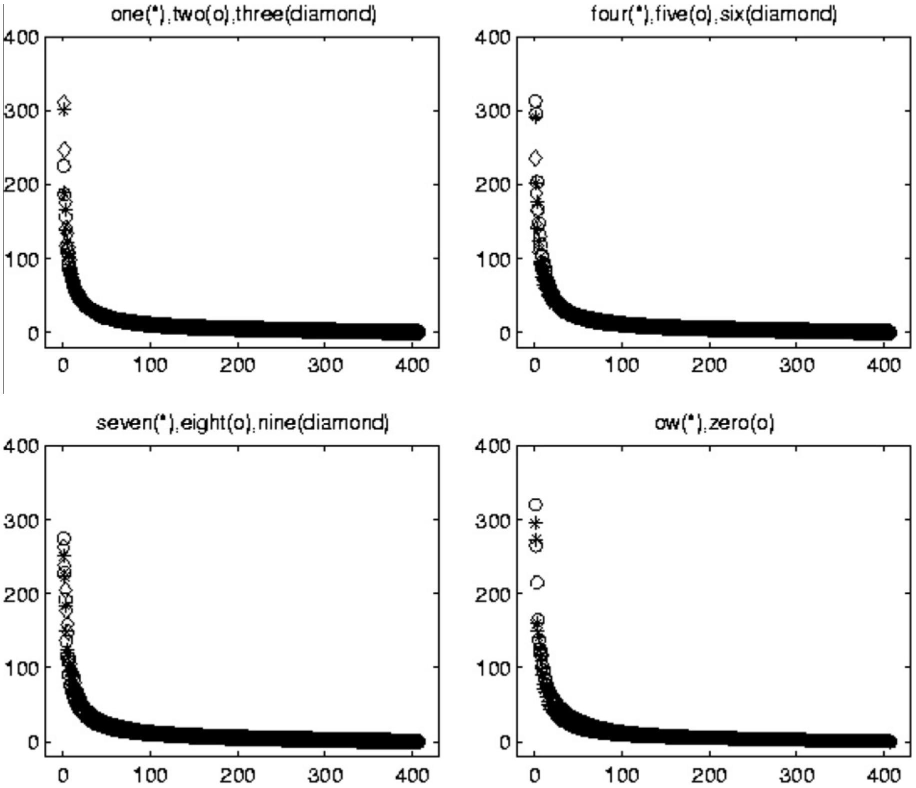

Fig. 1. The variations of the square roots of the eigenvalues of the within-class scatter matrices obtained for all digits.

sum of the two subspaces **B** and $\mathbf{B}^{\perp}$ is the whole space, and their intersection is the null space.

The eigenvalues of the covariance matrix $\mathbf{\Phi}$ are non-negative and they can be written in decreasing order: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ be the orthonormal eigenvectors corresponding to these eigenvalues. The first $(m-1)$ eigenvectors of the covariance matrix correspond to the nonzero eigenvalues. Therefore, the common vector can be shown as the linear combination of the eigenvectors corresponding to the zero eigenvalues of $\mathbf{\Phi}$ (Gülmezoğlu et al., 2001), that is,

$$\mathbf{a}_{com} = \langle \mathbf{a}_i, \mathbf{u}_m \rangle \mathbf{u}_m + \cdots + \langle \mathbf{a}_i, \mathbf{u}_p \rangle \mathbf{u}_p \quad \forall i = 1, 2, \ldots, m \tag{7}$$

From here, the common vector $\mathbf{a}_{com}$ is the projection of any feature vector onto the indifference subspace $\mathbf{B}^{\perp}$.

2.3.2. The sufficient data case

In this case, $m > p$. Let the difference subspace **B** be spanned by the orthonormal basis vectors $\mathbf{u_j} \in \mathbf{R}^p$ for $j = 1,2,\ldots,k-1$ $(k-1 < p)$, and let the indifference subspace $\mathbf{B}^{\perp}$ be spanned by the orthonormal basis vectors $\mathbf{u_j} \in \mathbf{R}^p$ for $j = k, k+1, \ldots, p$.

The value of k can be chosen in that the sum of the smallest eigenvalues is less than some fixed percentage L of the sum of the entire set (Gülmezoğlu, Dzhafarov, Edizkan, & Barkana, 2007; Oja, 1983). This value can also be determined from the point where the eigenvalues of the training data start to vary slowly when the eigenvalues are plotted in descending order as shown in Fig. 1 (Gülmezoğlu et al., 2007).

The orthogonal projection matrix **P** on the difference subspace **B** will be

$$\mathbf{P} = \sum_{j=1}^{k-1} \mathbf{u}_j \mathbf{u}_j^T \tag{8}$$

and the orthogonal projection matrix $\mathbf{P}^{\perp}$ onto the indifference subspace $\mathbf{B}^{\perp}$ will be

$$\mathbf{P}^{\perp} = \sum_{j=k}^{p} \mathbf{u}_j \mathbf{u}_j^T \tag{9}$$

P and $\mathbf{P}^{\perp}$ are symmetrical, idempotent $p \times p$ matrices, $\mathbf{P} + \mathbf{P}^{\perp} = \mathbf{I}$ where **I** is the identity matrix. The purpose in the decomposition of the whole feature space into two subspaces is to eliminate some part of the whole space containing large variations from the mean.

From here, the common vector $\mathbf{a}_{com}$ can be obtained as follows (Gülmezoğlu et al., 2007):

$$\mathbf{a}_{com} = \mathbf{P}^{\perp} \mathbf{a}_{ave} = \mathbf{P}^{\perp} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \right) \tag{10}$$

The projection of any feature vector in word-class c onto the indifference subspace will be close to the common vector of that class.

2.3.3. Decision criterion

For the recognition process, the following decision criterion is used:

$$K^* = \underset{1 \leqslant c \leqslant C}{argmin} \left\| \mathbf{P}^{c\perp} \left(\mathbf{a}_x - \mathbf{a}_{ave}^c \right) \right\|^2 \tag{11}$$

where $\mathbf{a}_x$ is the unknown test vector. If the distance is minimum for any class c, the unknown vector $\mathbf{a}_x$ is assigned to class c.

3. Experimental work

The TI-digit database is used in this experimental work. In our TI-digit database, there are 112 speakers repeating each digit twice in the TI-training set and 111 speakers again repeating each digit twice in the TI-test set. Therefore, the database includes 224 repetitions in the training set and 222 repetitions in the test set.

In the experimental work, raw speech data corresponding to each repetition is initially taken with silence regions. The center

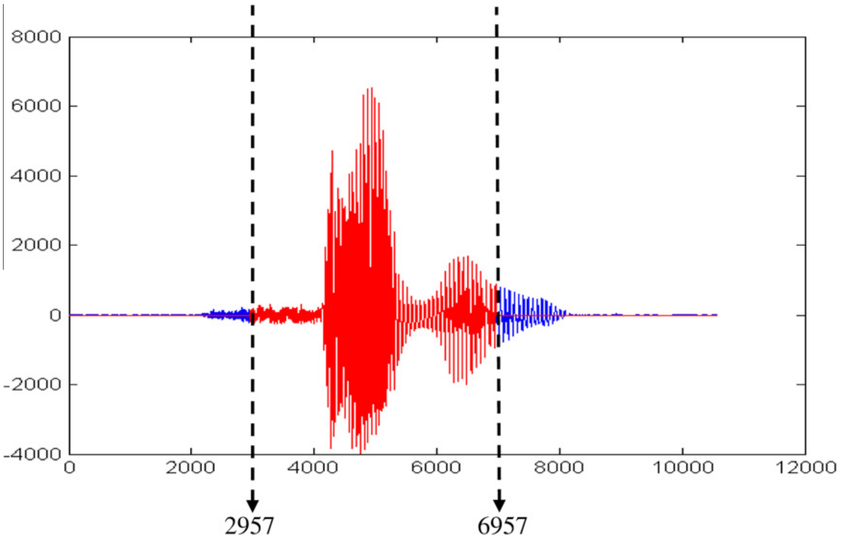

Fig. 2. The end points of the word "seven" when the COG formula (1) is used.

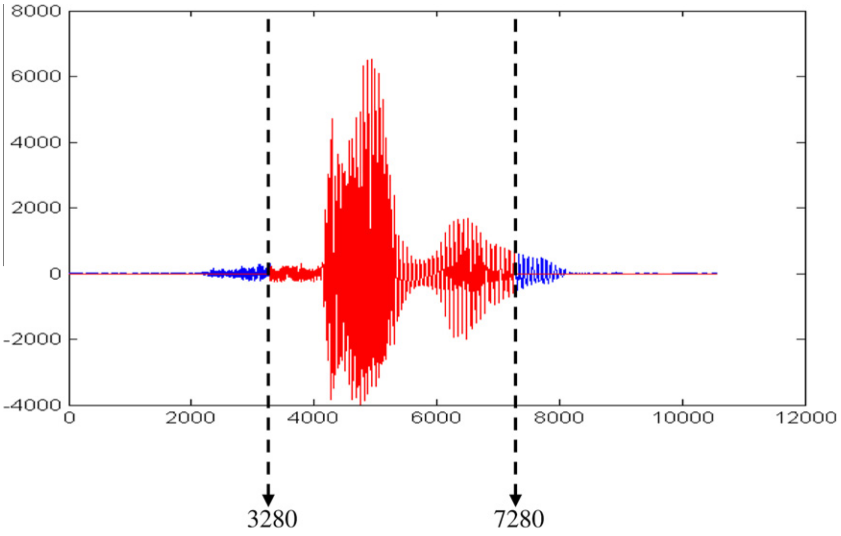

Fig. 3. The end points of the word "seven" when the COG formula (2) is used.

of gravity (η) of this data is found and then 2000 samples occurring before and after η are taken and stored in the computer. This number of samples is empirically determined from the training set of the database. Thus, each repetition has 4000 samples, as shown in Figs. 2–5. After pre-emphasizing all samples, each repetition is divided into eight frames and the Hamming window is applied to each frame. The overlap between the frames is set to a 1/4 of the number of samples in each frame. Thirty-three root-melcep parameters are calculated and stacked to construct the feature vector with the dimension of 330×1 for each repetition of each digit (Ergin, Gülmezoğlu, & Barkana, 2004). In order to investigate the effect of sample size on the performance of CVA classifier, the different number of samples (1500 and 2500) occurring before and after η are also used in the experimental study.

If the endpoint detection algorithm is used in our TI-digit database, the algorithm gives reasonable endpoint locations, as can be seen in Figs. 6 and 7. Following the end-point detection, 330 root-melcep parameters, calculated by the same feature extraction method previously mentioned, are used to represent the isolated words.

3.1. The insufficient data case

Since the dimension of the feature vectors is 330, 224 feature vectors in the training set are used in the calculation of the covariance matrices for the insufficient data case. Therefore, the covariance matrix (330×330) of each class has 107 (=330 − 224 + 1) zero eigenvalues. The projection matrix is constructed by considering the eigenvectors corresponding to zero eigenvalues. Two hundred and twenty-two feature vectors in the test set are used to evaluate the performance of CVA classifier. When the decision criterion in Eq. (11) is employed, the average recognition rates of all digits for the training and test sets are given in Table 1 for three different number of samples (3000, 4000 and 5000). The results obtained from the end-point detection algorithm are also given in Table 1.

3.2. The sufficient data case

In this case, in order to satisfy the condition ($m > p$), 426 feature vectors out of 446 (224 from TI training set and 222 from TI test

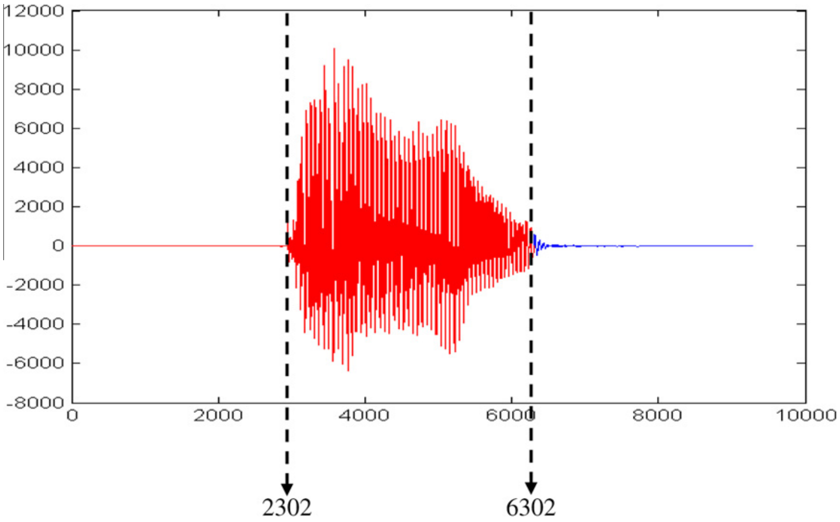

Fig. 4. The end points of the word "ow" when the COG formula (1) is used.

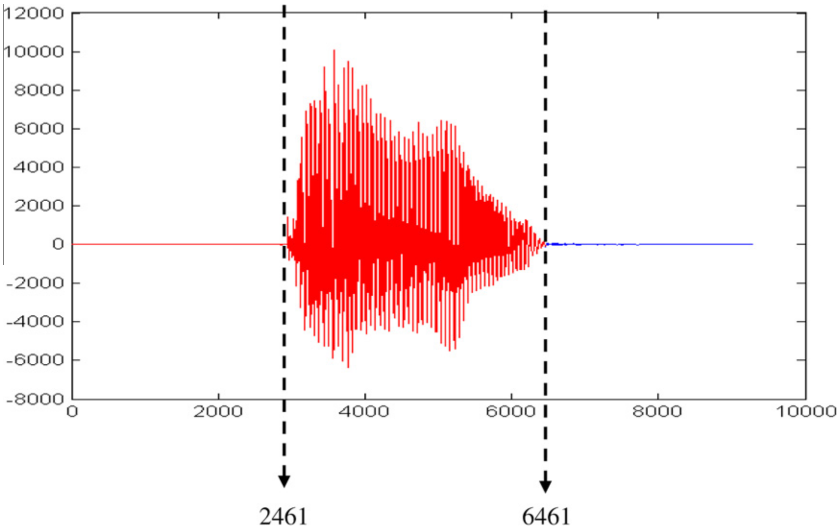

Fig. 5. The end points of the word "ow" when the COG formula (2) is used.

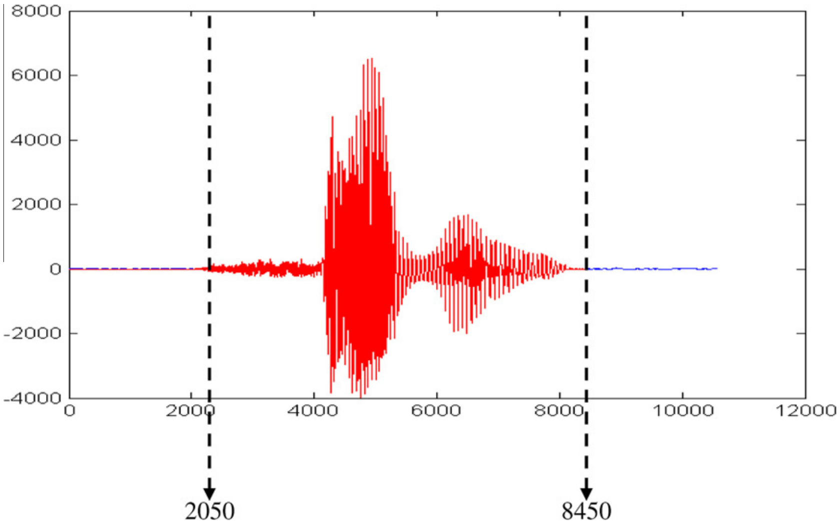

Fig. 6. The end points of the word "seven" when the endpoint detection algorithm is used.

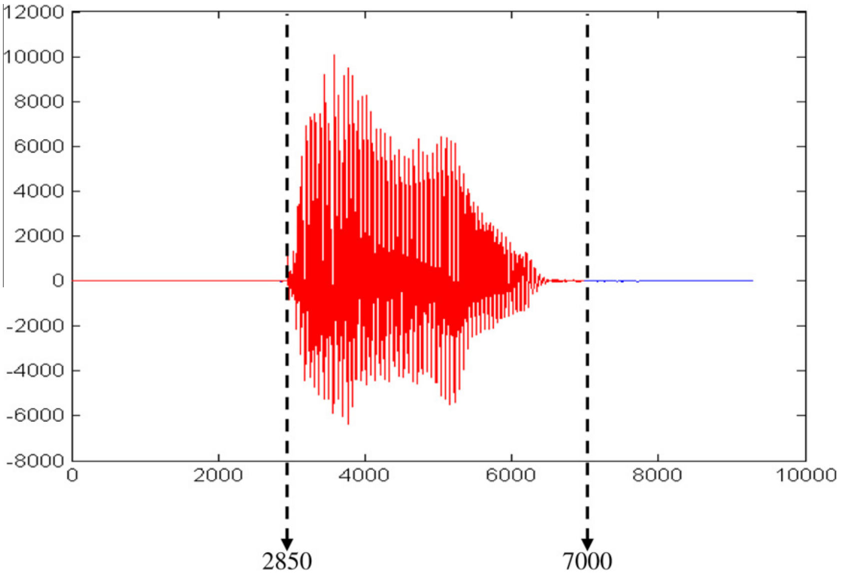

Fig. 7. The end points of the word ‘‘ow” when the endpoint detection algorithm is used.

Table 1
Average recognition rates (%) of 11 digits.

Methods	Cases	Sample range	Training	Testing
COG formula (1)	Insufficient case	±1500	100	95.17
		±2000	100	95.50
		±2500	100	95.58
	Sufficient case	±1500	99.93	97.19
		±2000	99.94	98.39
		±2500	99.93	98.60
COG formula (2)	Insufficient case	±1500	100	94.72
		±2000	100	95.66
		±2500	100	95.50
	Sufficient case	±1500	99.86	97.11
		±2000	99.89	98.14
		±2500	99.92	98.60
End point detection	Insufficient case	None	100	93.57
	Sufficient case		99.11	97.36

set) are used in our new training set and the remaining 20 feature vectors are used in our new test set. Therefore, the feature vectors in the training and test sets are completely disjoint in the experiments. An equal number of male and female speakers is used in the training set, as well as in the test set.

The dimension (k) of the indifference subspace is based on the selection of L. In our experimental work, L is taken as 5%, at which a good performance is obtained while retaining a small proportion of the variance present in the original space (Gülmezoğlu et al., 2007; Swets & Weng, 1996). $L = 5\%$ for indifference subspace was attained at a different number of eigenvalues for each digit. The average number of these eigenvalues is equal to 255.

Two COG formulas are used in the sufficient data case. Since 20 feature vectors in the test set are too few to determine recognition accuracy, the ‘‘leave-20-out” method is applied. Thus, the testing process is repeated 11 times to cover all the repetitions in the TI test set. The average recognition rates obtained from these iterations are given in Table 1. It must be pointed out that our training set can be completely classified with just a few eigenvalues. For example, when the eigenvectors corresponding to the smallest 10 eigenvalues are used, a 100% recognition rate is obtained. When the end-point detection algorithm is used in the sufficient data case, the results obtained from the ‘‘leave-20-out” method are also given in Table 1.

4. Conclusion

In this paper, the COG method is compared with one of the classical endpoint detection methods in view of recognition performance obtained from the CVA method. The COG method determines the point around which the momentum through the word is balanced. If the preceding and succeeding regions around the COG point change for different classes, the recognition can be almost perfect. The reason of this is that the COG approximately gives the same regions in all words of each class so that it improves the CVA model parameters and prevents mismatch between the training and test sets. Also CVA takes the invariant features of the word with all the varying features being eliminated in the indifference subspace. More specifically, the features that are obtained in the first few frames and in the last few frames would vary depending on the lengths of the word. Consequently, these features will be eliminated before the recognition process.

All the experimental studies are carried out for isolated words in the TI-digit database under noise-free environment. As seen from the Table 1, the results obtained from the COG method with CVA are higher than those obtained from the classical endpoint detection with CVA. Therefore, COG with CVA is suitable for isolated word recognition with limited number of words. The average recognition rates of the test set obtained with the COG formulas slightly change for the different number of samples. However, different number of samples occurring before and after center of gravity can be selected especially for short words as in the TI-Digit database.

The authors are aware that with the COG method, some of the initial and final frames of long words will not be taken into account in the recognition process. The recognition accuracy will decrease for words which are long enough to have important information outside the analysis window. It must be noted that the words with similar phonemes in their centers can be misclassified. Furthermore, some of the initial and final frames of short words will belong to the silence regions.

By taking a suitable number of samples from the left and right of the center of gravity, almost all speech data can be extracted from any recording. It has also been shown that this speech data is satisfactorily recognized by the CVA classifier. The proposed COG method with the CVA can be efficiently used in real time recognition of isolated words.

If the endpoint detection algorithms do not detect the real endpoints, the recognition rates of CVA may deteriorate. However, the COG method will almost never fail in finding the point of the center of gravity in the isolated word, although it never provides the real endpoints. Moreover, with the COG method, the recognition rates of the CVA will not deteriorate at all.

References

Çevikalp, H., Neamtu, M., Wilkes, M., & Barkana, A. (2005). Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1–10.

Ergin, S., Gülmezoğlu, M. B., & Barkana, A. (2004). Use of improved feature vectors in affine transformation for robust speech recognition. In *Proceedings of international conference on information*.

Feth, L. L., Fox, R. A., Jacewicz, E., & Iyer, N. (2002). Dynamic center-of-gravity effects in consonant–vowel transitions. In *Dynamics of speech production and perception, NATO advanced study institute* (pp. 5–13).

Gülmezoğlu, M. B., Dzhafarov, V., Keskin, M., & Barkana, A. (1999). A novel approach to isolated word recognition. *IEEE Transactions on Acoustic Speech and Signal Processing, 7*(6), 620–628.

Gülmezoğlu, M. B., Dzhafarov, V., & Barkana, A. (2001). The common vector approach and its relation to principal component analysis. *IEEE Transactions on Speech and Audio Processing, 9*(6), 655–662.

Gülmezoğlu, M. B., Dzhafarov, V., Edizkan, R., & Barkana, A. (2007). The common vector approach and its comparison with other subspace methods in case of sufficient data. *Computer Speech and Language, 21*, 266–281.

Günal, S., Edizkan, R., & Barkana, A. (2003). The design of real-time digit recognizer using the common vector approach. In *Proceedings of 11th national conference on signal processing and applications* (pp. 308–311).

Huang, L., & Yang, C. (2000). A novel approach to robust speech endpoint detection in car environments. In *Proceedings of international conference on acoustics, speech, and signal processing* (pp. 1751–1754).

Landi, G. (2003). Properties of the center of gravity algorithm. In: *Proceedings of 8th international conference on advanced technology and particle physics*.

Li, Q., Zheng, J., Tsai, A., & Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing, 10*, 146–157.

Oja, E. (1983). *Subspace methods of pattern recognition*. New York: John Wiley & Sons Inc.

Orlandi, M., Santarelli, A., & Falavigna, D. (2003). Maximum likelihood endpoint detection with time-domain features. In *Proceedings of European conference on speech communication and technology* (pp. 1757–1760).

Orság, F., & Zbořil, F. (2003). Endpoint detection in the continuous speech using the neural networks. In *Proceedings of 37th international conference on modelling and simulation of systems* (pp. 7–13).

Rabiner, L. R., & Sambur, M. R. (1975). An algorithm for determining the endpoints for isolated utterances. *The Bell System Technical Journal, 54*, 297–315.

Rabiner, L. R. (1978). *Digital processing of speech signals* (1st ed.). New Jersey: Prentice Hall.

Raj, B., & Singh, R. (2003). Classifier-based non-linear projection for adaptive endpointing of continuous speech. *Computer Speech and Language, 17*, 5–26.

Shen, J., Hung, J., & Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In *Proceedings of international conference on spoken language processing*.

Shin, W., Lee, B., Lee, Y., & Lee, J. (2000). Speech/non-speech classification using multiple features for robust endpoint detection. In *Proceedings of international conference on acoustics, speech, and signal processing* (pp. 1399–1402).

Stylianou, Y. (1998). Removing phase mismatches in concatenative speech synthesis. In: *Proceedings of 3rd ESCA speech synthesis workshop* (pp. 267–272).

Stylianou, Y. (1999). Synchronization of speech frames based on phase data with application to concatenative speech synthesis. In *Proceedings of 6th European conference on speech communication and technology* (pp. 2343–2346).

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing, 9*, 21–29.

Swets, D. L., & Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*, 831–836.

Van Son, R. J. J. H., & Pols, L. C. W. (1996). An acoustic profile of consonant reduction. In *Proceedings of 4th international conference on spoken language processing* (pp. 1529–1532).

Wu, G., & Lin, C. (2000). Word boundary detection with mel-scale frequency bank in noisy environment. *IEEE Transactions on Speech and Audio Processing, 8*, 541–554.