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.

© 2010 Elsevier Ltd. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/386029

Download Persian Version:

https://daneshyari.com/article/386029

Daneshyari.com