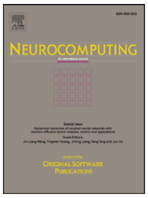
Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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



A weighted accent classification using multiple words

Muhammad Rizwan*, David V. Anderson

Georgia Institute of Technology, Atlanta, GA, U.S.A.

ARTICLE INFO

Article history:
Received 2 June 2016
Revised 11 January 2017
Accepted 13 January 2017
Available online xxx

MSC:
00-01
99-00

Keywords:
Extreme learning machines
Support vector machines
Accent classification

ABSTRACT

Speech recognition systems exhibit performance degradation due to variability in speech caused by the accents or dialects of speakers. This can be overcome by correctly identifying the accent or dialect of the speaker and using accent or dialect information to adapt speech recognition systems. In this paper, we apply extreme learning machines (ELMs) and support vector machines (SVMs) to the problem of accent/dialect classification on the TIMIT dataset. We used Mel frequency cepstrum coefficients (MFCCs) and the normalized energy parameter along with their first and second derivatives as raw features for training ELMs and SVMs. A weighted accent classification algorithm is proposed that uses a novel architecture to classify North American accents into seven groups. Using this algorithm, we obtained a classification accuracy of 77.88% using ELMs, which to our knowledge, is the best result reported for accent classification on the TIMIT dataset. We also compared the performance of ELMs with SVMs as classifiers for our weighted accent classification algorithm and with multi-class classification using ELMs or SVMs.

1. Introduction

Speech signals intrinsically exhibit many variations, even in the absence of background noise. The three most prominent types of variations are due to acoustic effects, accent, and dialect. Acoustic variations are primarily related to inherited physical characteristics of size and shape of a vocal tract. Two different people saying the same sentence results in different spectrograms. The variations due to accent result from the relative prominence of a particular syllable or a word in pronunciation determined by the regional or social background of the speaker [1]. Different accents effect a change in the order and number of phonemes used to construct each word of an utterance, i.e., phoneme deletion, insertion and substitution with respect to some reference accent. Dialect is defined as a regional variety of a language distinguished by pronunciation, grammar or vocabulary. Every individual develops a characteristic speaking style at an early age that depends heavily on his or her language environment as well as the region where the language is spoken [2,3]. We are using a database of read speech that is labeled by dialect region. However, since the speech is read, although we are using dialect regions, words and grammar choices are eliminated, leaving only accent variations. Accordingly, in this paper we will use these terms interchangeably.

Lawson, et al. showed by their cross accented experiments that phonemic models obtained from a different accent were 1.8 times less accurate in recognizing speech than those from a matched accent [4]. Performance of a speech recognizer can be further improved by adapting a system based on accent/dialect. Goronzy achieved a 37% reduction in word error rate (WER) by adapting a recognizer based on accent [5]. There has been little past research in the area of accent classification. In particular, most of the previous work in the field involves accent classification among non-native English speakers. Accent variation among native American speakers is more challenging and has not enjoyed the same amount of attention in speech community research.

Choueiter, et al. extended language identification techniques to a large-scale accent classification task [6]. They performed several experiments using heteroscedastic linear discriminant analysis (HLDA) and maximum mutual information (MMI) on the Foreign Accented English (FAE) dataset [7]. The FAE is composed of utterances spoken by native speakers of 23 languages. They found that acoustic-only methods are quite effective for accent classification in contrast to typical language identification systems. Angkititrakul and Hansen used a phoneme-based model to design a text independent automatic accent classification system [8]. They performed experiments capturing the spectral evolution information as potential accent sensitive cues. They generated subspace representations using principal component analysis (PCA) and linear discriminant analysis (LDA). They compared a spectral trajectory model framework with a traditional hidden-Markov-model (HMM) recognition framework using an accent sensitive word corpus. Sys-

* Corresponding author.
E-mail address: mrizwan@gatech.edu (M. Rizwan).

http://dx.doi.org/10.1016/j.neucom.2017.01.116
0925-2312/© 2017 Elsevier B.V. All rights reserved.

tem evaluation was performed using a corpus that represent five English speaker groups, which consisted of native American English and English speakers having Mandarin Chinese, French, Thai, and Turkish accents for both male and female speakers. Guarasa used Gaussian mixture models (GMMs) and Bayes' classifiers for German versus Spanish accent classification [9]. Clopper, et al. did an extensive study of vowel variation among different regions of North America by acoustic measures of duration and first and second formant frequencies [10]. Hansen, et al. did an extensive analysis and modeling of speech under accents on NATO N-4, TIMIT and the WSJ corpus [11]. They analyzed prosodic structure (formants, syllable rate and sentence duration), phoneme acoustic space and did word-level based modeling on large vocabulary data. In their experiments, they found that using the most discriminating vowels from each group improves the accent detection rate.

In this paper, we propose an accent classification algorithm based on extreme learning machines (ELMs). ELMs are attractive for the accent classification task as they can be quickly trained and also provide a better generalization capability for small amounts of training data [12,13]. We also compare our accent classification algorithm performance by using support vector machines as classifiers. The rest of the paper is organized as follows: the theory of extreme learning machines (ELMs) and support vector machines (SVMs) is presented in Sections 2 and 3 with a comparison between ELMs and SVMs in Section 4. In Section 5, we propose our accent classification algorithm. A description of experiments performed, including the dataset, extraction of features, ELM training, and SVM training for the weighted accent classification algorithm is presented in Section 6. Results are discussed in Section 7 and the conclusion, with suggestions for future work presented in Section 8.

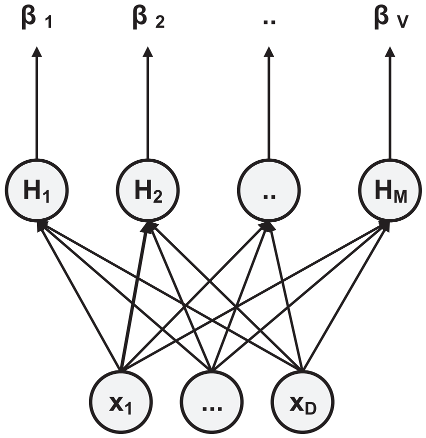


Fig. 1. Extreme learning machines.

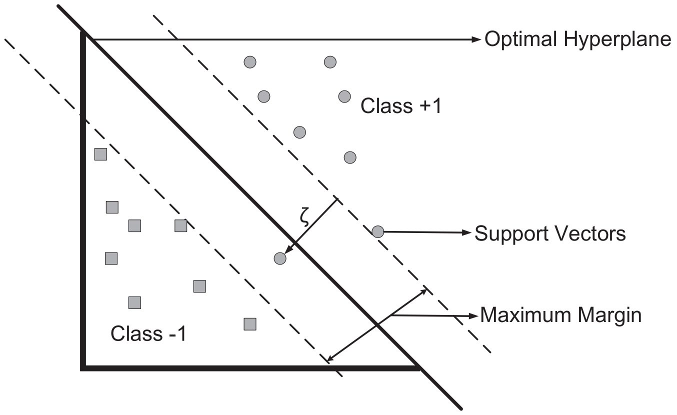


Fig. 2. Support vector machine.

2. Extreme learning machines

Extreme learning machine (ELM) is a robust learning algorithm for single layer feed-forward neural networks (SLFNs) [14]. Currently, SLFNs mostly use gradient based methods for training neural networks. Gradient based methods often get trapped in local minima and, as a result, give suboptimal solutions. Genetic and evolutionary algorithms have also been used to overcome local minima problems, but they are computationally expensive [15].

In ELMs, input weights of the hidden layer neurons are randomly generated and output weights of the hidden layer neurons are learned analytically [14,16]. By learning weights analytically, there is a great performance speedup for training neural networks as compared with learning methods such as back-propagation [17]. Theoretically, it has been shown that by using ELMs universal approximation can be achieved [18,19]. ELMs can also be used for training multilayer perceptrons by using hierarchical frameworks [20].

Various other architectures for ELMs have been proposed. In incremental-ELM, hidden nodes are added incrementally and output weights are determined analytically [21]. In online sequential-ELM, training data is fed to the network in chunks [22]. Local receptive fields-ELM uses local structures and combinatorial nodes for incorporating translational invariance in the network [23]. ELMs can be used for both regression and multiclass classification problems directly [24].

ELMs transform the input data to the hidden layer by via randomly initialized weighted connections. A single hidden layer network with M hidden nodes is shown in Fig. 1.

The output function of the single layer network with M hidden neurons can be written as [12]:

$$f(\mathbf{x}) = \sum_{i=1}^{M} \boldsymbol{\beta}_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \tag{1}$$

where

$$h_i(\mathbf{x}) = \sigma(\mathbf{w_i x} + b_i) \tag{2}$$

and σ is a non-linear activation function given by:

$$\sigma(\mathbf{w}_i\mathbf{x} + b_i) = \frac{1}{1 + e^{-(\mathbf{w}_i\mathbf{x}+b_i)}} \tag{3}$$

$\boldsymbol{\beta}$ is the vector of weights between M neurons in the hidden layer and the output layer:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} \tag{4}$$

The goal of ELM is to minimize the training error as well as the norm of the output weights. It does not require any adjustments to the input weights of neurons Fig. 2 in the hidden layer [12,23,25,26].

$$\text{minimize} \quad \left\|\boldsymbol{\beta}\right\|_p^{\sigma_1} + C\left\|\mathbf{H}\boldsymbol{\beta} - \boldsymbol{T}\right\|_q^{\sigma_2} \tag{5}$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, and $p, q = 0, 1, 2, \ldots, \infty$, and $\mathbf{H}$ is the output matrix at the hidden layer given by:

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \vdots \\ \mathbf{h}(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} h_1(\mathbf{x}_1) & \ldots & h_M(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ h_1(\mathbf{x}_N) & \ldots & h_M(\mathbf{x}_N) \end{bmatrix}. \tag{6}$$

Please cite this article as: M. Rizwan, D.V. Anderson, A weighted accent classification using multiple words, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.01.116