International Journal of Soft Computing 2 (2): 335-338, 2007

Stressed/Neutral Speech Classification Using Gaussian Support Vector Machines

[1]T. Santhanam [2]M. Nachamai [3]M. Muthuraman and [4]C.P. Sumathi
[1]Department of Computer Science PG and Research D.G. Vaishnav College,
Arumbakkam, Chennai-106, India
[2]Department of Computer Science Alliance Business Academy, Bangalore-76
[3]Institute of Circuits and Systems University of Kiel, Germany-24108
[4]Department of Computer Science SDNB Vaishnav College for Women,
Chromepet, Chennai-44, India

Abstract: Artificial Neural Networks (ANN) is one of the approaches used in acoustic modeling. ANNs are better suited for handling complicated tasks. They are used to recognize speech and also to handle even low quality, noisy and speaker independent data with an improved efficiency i.e., ANN has displaced the most frequently employed Hidden Markov Model (HMM) for speech prob lems in all aspects in general and scalability in particular. Scalability in ANN is very less compared with HMM when provided with huge a mount of training data. This study makes use of ANN for classification of phonemes and style (stressed speech or neutral speech). Better results are obtained by considering only the pitch contour, one among the many base features of speech, as input to the network than giving multiple inputs. The Gaussian-kernel estimator function capable of mapping the data into a high dimensional space takes care of the scalability feature of ANN. This approach has resulted in enhanced recognition rates of speaking style.

Key words: HMM, ANN, G-SVM, MLNN, PCA

INTRODUCTION

Speech recognition is the process of converting a speech signal to a set of words, by means of an algorithm implemented as a com puter program. Speech recognition applications that have emerged over the last years include voice dialing (e.g., Call home), call routing (e.g., I would like to make a collect call), simple data entry (e.g., entering a credit card number) and preparation of structured documents (e.g., a radiology report). Voice or speaker recognition is a related process that attempts to identify the person speaking, as opposed to what is being said. Speech recognition is a difficult problem, largely because of the many sources of variability associated with the signal, namely.

Phonetic variabilites: The acoustic realizations of phonemes, the smallest sound units of which words are composed, are highly dependent on the context in which they appear.

Acoustic variabilities: Result from changes in the environment as well as in the position and characteristics of the transducer.

Speaker variabilities: Result changes in the speaker's physical and emotional state, speaking rate, voice quality.

Hidden Markov Model is the underlying technology for speech recognition. Researchers have also adopted dynamic algorithms, neural networks and knowledge-based approaches extensively in speech recognition.

ANNs are capable of solving much more complicated recognition tasks, but do not scale like HMMs for large vocabularies. Apart from being used in general-purpose speech recognition applications they can also handle low quality, noisy data and speaker independence. They can achieve greater accuracy than HMM based systems, as long as there is data for training with limited vocabulary. A more general approach using neural networks is phoneme recognition. This is an active field of research and generally the results are far better than that of HMMs. There is also NN-HMM hybrid systems that use the neural network part for phoneme recognition and the hidden markov model part for language modeling. ANN can outperform the HMM with minimal input.

A novel approach employed in this study is presented below.

Step 1: Neural Predictive Coding (NPC) network is used for phonetic classification because NPC can account for the non-linear features of the speech input, rather than the linear techniques like Linear Predictive Coding (LPC) and Perceptual Linear Predictive analysis (PLP).

Corresponding Author: T. Santhanam, Department of Computer Science PG and Research D.G. Vaishnav College, Arumbakkam, Chennai -106, India

Step 2: The probability of the phonetic classes are calculated using Maximum Likelihood Neural Network (MLNN) that uses the Likelihood function on the acoustic occurrence and not the minimization of mean error employed by the other standard methods, for getting a better acoustic score.

Step 3: Ergodic scoring method of HMM is applied to get a better measure of accuracy for detecting the vowel nucleus of words, which is the major element of lexical stress.

Step 4: Pitch contour (F0), one of the prosodic features is chosen to increase the computational efficiency of the ANN.

Step 5: GSVM that applies the kernel-trick to scale into a high-dimensional feature space for handling larger vocabularies is used for classification.

The study is organized as follows. The next section explains the technical aspects and the algorithm used in the research. Experiments conducted and results are presented in section three followed by conclusion in the last section.

MATERIALS AND METHODS

The method adopted in this research accounts for the robustness in speech and low quality inputs. Figure 1 portrays the flowchart of this method for pattern recognition in speech. The three steps involved in the algorithm are phonetic classification phase, vowel element recognition phase and. speaking-style recognition phase.

Phonetic classification: The first phase classifies the incoming speech signal into phonetic classes. Feature extraction in speech is generally carried out using temporal methods like LPC and PLP, which are data driven approaches by nature. Neural networks can achieve an extension of LPC to non-linear domain that is called the Neural Predictive Coding (NPC). Due to the pronounced non-linear features of the speech signal, NPC method is adopted for phonetic classification. The objective of applying the NPC on the incoming speech signal (speaker independent) is to extract the phonetic features in a non-linear plane (Chetouani *et al*., 2004.) . The NPC comprises of two sets of weights (referred to as codes) where the common parts of the speech production model are given to the first layer as inputs (input layer to the hidden layer) and specific parts are given as inputs to the second layer (hidden layer to the output layer). The former is called the Parameterization Phase (PP) and the latter is called the Coding Phase (CP) Fig. 2.

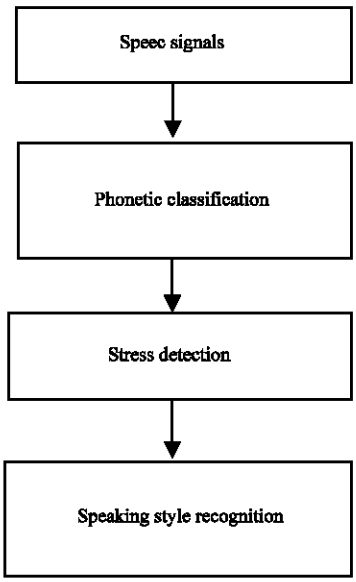

Fig. 1: Flowchart showing the stages of speech recognition

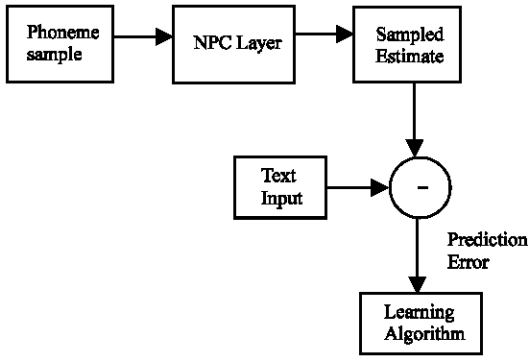

Fig. 2: Neural predictive coding architecture

The CP generates the coding vector, which carries discriminant phonetic features for each class. Let i and j be the two phonemes belonging to two different classes C_i and C_j. The NPC models for the two phonemes are,

$$F_{w,a_i} = H_{a_i} \times G_w$$
$$F_{w,a_j} = H_{a_j} \times G_w$$

where H_{a_i} and H_{a_j} are distinct functions for the two different classes and G_w depicts the common features in PP. Once the phonetic units of speech are identified, the probability of each phonetic unit is computed using the Maximum Likelihood Neural Network (MLNN). An independent feature extraction method used as the optimization criterion may cause inconsistency between the feature extraction and classification of the recognizer

resulting in degradation of performance of the classifiers. Thus, MLNN is used to increase the efficiency of the method to research with input from different environments (Juang and Katagiri *et al*., 1992). Neural networks are generally applied to minimize the Mean Standard Error (MSE)

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(t_i - O_i)$$

where n-number of nodes, t-target output and O-observed/acquired output. There is no guarantee that minimizing MSE will maximize the acoustic scores. MLNN does not minimize the MSE; rather it calculates the maximum likelihood of the acoustic sequence (Yuk *et al*., 1999). The conditional probability of the next word to come in sequence with the first word being fixed is calculated by MLNN, thus increasing the probabilistic score of the occurrence.

Stress detection: The next phase is vowel element recognition in which the identification of stress is carried out. The vowel element in a word contributes more towards the stress of the word. In order to determine the lexical stress, the vowel nucleus in the word is found by word spotting. Word spotting is calculated using the Ergodic-HMM (E-HMM), which shows better results based on recognition rate, word-end detection and mean-hypothesis length (Ozeki, 1996). In this method, for each frame t and state j of the E-HMM, the accumulated log probability Le (t, j) is computed using Viterbi search. Then combining the accumulated log-likelihood L (t,Wm, i) and Le (t, j) another score for Wm terminating at frame t is defined and calculated as,

$$S_e(t, W_m) = \{L(t, W_m, s_m) - L(t, W_m \cdot, s_m \cdot)\} - \{MaxL_e(t, j) - MaxL_e(t', j)\}$$

A number of features like MFCC, log- energy, pitch, duration and speech-rate may be considered for stress detection but in the current research the emphasis is only on pitch feature since pitch feature at word level seem to resolve to a better efficiency than combining multiple features in detection of stress. The pitch contour (F0) can easily distinguish the emotional states and thus can identify between stressed speech and neutral speech with better accuracy. From the pitch contour the following eight features for each vowel segment are extracted:

- Mean pitch (normalized)-mean pitch value of the vowel normalized by the mean pitch of the entire utterance.
- Pitch value at start point (normalized)-the pitch value at the start point of the vowel divided by the mean pitch of the entire utterance.
- Pitch value at end point (normalized)-the pitch value at the end point of the vowel divided by the mean pitch of the entire utterance.
- Maximum pitch value (normalized)-maximum pitch value of the vowel divided by the mean pitch of the entire utterance.
- Minimum pitch value (normalized)-minimum pitch value of the vowel divided by the mean pitch of the entire utterance.
- Relative pitch difference-difference between the normalized maximum and minimum pitch values (inference to be made-a negative value indicates a falling pitch and a positive value denotes a rising pitch).
- Absolute difference- magnitude of the relative difference.
- Pitch trend-sign of the relative difference-value 1 if the pitch rises over the vowel segment and a value-1 if it falls and a value of 0 if it is flat.

For normalizing the input pitch feature space, Principal Component Analysis (PCA) is used which gives the projected feature space. The PCA transforms the given data set X of dimension M to an alternative data set Y of a smaller dimension L using the co-variance method.

$$Y = KLtransform(X)$$

Speaking style recognition: The final phase recognizes and classifies the utterances into a neutral and stressed speech. A neural network is made use of to estimate class probabilities and the discriminate is more because the learning rule minimizes the classification error and maximizes the distinction between the classes. Fewer parameters are required for a neural network than the general models and is computationally most efficient. The general SVM uses the parametric space and hence cannot be used as a classifier in this research (Wang and Paluiwal, 2002). The GSVM employed here is a non-linear, supervised learning method that uses the kernel-trick to apply the linear classification techniques to non-linear classification problems. The kernel-trick is used to map the original observations into a higher dimensional non-linear space so that the linear classification in the new space is same as the non-linear classification in the original space. Gaussian-kernel function given below is applied that can generalize the test data beyond the limits

$$f = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{u^2}{2}}$$

$$u = (x - x_i)/h$$

where, h-bandwidth of the utterance, x_i-values of independent variables in the data (eight pitch values), x-value of the independent variable for which one seeks an estimate (F0-pitch contour). The function is unbounded on *x*-so every data point will be brought into the estimate.

GSVM can classify the classes, which are on multi-datasets and not linearly separable in the original space that are transformed into a high dimension feature space. There are only two classes (W_1, W2) in this research and the training data are labeled by the following rule:

$$Y_i = \begin{Bmatrix} +1 & \text{if } X_i \text{ belongs to class } W_1 \\ -1 & \text{if } X_i \text{ belongs to class } W_2 \end{Bmatrix}$$

RESULTS

Experiments were conducted on Speech Under Simulated Actual Stress (SUSAS) database created by the Robust Speech Processing Laboratory. The database is divided into four domains namely neutral, angry, lombard and loud encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), in the age group of 22 to 76 were employed to generate more than 16,000 utterances. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8 kHz. The isolated words employed in this research consist of three stressed classes (angry, loud and lombard) and the neutral class. Support Vector Machine (GSVM) used for speech classification is implemented in MATLAB. The results are compared with HMM (Kwon *et al*., 2003). In Table 1.

It has been reported in (Kwon *et al*., 2003). that the average accuracy of HMM was 96.3% and that of the proposed GVSM is 98.2%. The tabulated results and the average accuracy very clearly prove that the performance of GSVM is far better than HMM.

Table 1: Confusion matrix of the results obtained using the proposed GSVM and HMM (Kwon *et al*., 2003)

GSVM	Neutral	Stressed	HMM	Neutral	Stressed
Neutral	0.998	0.002	Neutral	0.918	0.082
Stressed	0.001	0.991	Stressed	0.023	0.977

CONCLUSION

Generally, neural networks operate successfully only with small sentences and with larger ones the accuracy drops down resulting in over fitting-data criterion; that is the network suffers with enormous amount of training data given, which either stops the network or degrades its performance. Most of the research research deals with the use of HMM for speech recognition. This study explored the possibility of using GSVM as an alternative to HMM to classify speech. The excellent outcome demonstrates the superiority of GSVM over HMM for speech recognition.

REFERENCES

Chetouani, M., B. Gas and J.L. Zarade, 2004. Learning Vector Quantization and Neural Predictive Coding for Nonlinear Speech Feature Extraction. Study presented at the XII Signal Processing Conference, Eusipco.

Juang, B.H. and Katagiri S. Discriminative, 1992. Learning for Minimum Error Classification. IEEE Transactions on Signal Processing pp: 40.

Kwon, Oh-Woof, Kwokleung Chan, Jiuchang Hao and Te-Won Lee, 2003. Emotion Recognition by Speech Signals, pp: 125-128.

Ozeki, Kazuhiko. 1996. Likelihood Normalization Using an Ergodic HMM for Continuous Speech Recognition. and ICSLP, 4: 2301-2304.

Wang, X. and K.K. Paliwal, 2002. Feature extraction for Integrated Pattern Recognition Systems. Fourth australian workshop on signal processing and applications, pp: 85-88.

Yuk, Dong Suk and James Flanagan, 1999. Telephone speech recognition using neural networks and hidden markov models. IEEE international conference on acoustics, speech and signal Processing, 1: 157-160.