Radial-Basis Functions Neural Network for Text Independent Speaker Recognition

A.A.Yakovenko, G.F.Malyhina,
Institute of Information Technology and Control Systems
St. Petersburg State Polytechnical University
St. Petersburg, Russia
e-mail: g_f_malychina@mail.ru
annother_@hotmail.com

Abstract — RBF neural network is proposed for solution the problem of text-independent speaker recognition. Recognition is based on estimation of sufficiently large set of acoustic features, construction of multidimensional histograms and approximation histograms with probability density functions with possibility of wide shape variation. Method allowed to reduce the probability of errors when decision was making.
Keywords—text independent identification, speaker recognition, radial-basis functions neural network.

1. INTRODUCTION

Biometric recognition systems allows us to find the right connection to authorize any person in information systems. In recent years the interest in voice biometrics has been increased [4, 5]. This is completely in demand in the areas of organization access permissions in information systems, biometric solving search and forensic accounting, voice verification of the driver and passengers, in the management elements of smart home, in banking systems, contact centers, etc. Identification of speaker's voice provides a unique opportunity to secure access to information, remote maintenance and examination to establish the identity.

Speaker identification and speaker verification problem is divided into two tasks a text-dependent identification and text-independent identification and can run using open set of speakers or closed set [4]. In the case of a closed set of speakers, phonogram will obviously belong to a particular individual, but if the phonogram does not belong to any candidate, then the problem is solved on an open set of speakers.

If identification system trained in advance to recognize universe passphrase delivered by announcer, then it is a text-dependent identification system. Phonemic dictionary and phrase structure in this case requires smaller amount of training speech data. The necessity of pronouncing passphrase during training and during the operation of the system limits the practical range of its application.

Identification system based on text-independent approach does not contain information about the uttered phrase. It is trained and then tested on arbitrary voice and speech data. The effectiveness of such identification systems is lower than in the text-dependent. But voice recognition in this case has broader application, since knowledge of uttered phrase is optional.

Voice identification reduces to the problem of deciding which of the plurality of speakers most likely belongs to the tested track. Since human speech is regarded as an acoustic signal, the analysis of the signal takes place by means of digital processing.

Develop a system of identification occurs in three stages [3]: on the first stage implementation of features extraction, on the second - modeling of speakers and on the third stage - decision-making is carried out. Thus, in general, a standard system for speaker voice recognition extracting unit comprise primary feature vectors of the speech signal and the simulation unit speaker's voice, which are divided according to the tasks. Since actual recordings made under conditions, there are many extraneous signals, various kinds of noise, impulse noise and congested areas of speech, preprocessing and noise removing stage can improve the efficiency further processing.

Special pre-processing algorithms of the entire signal, perform the selection of speech segments, and feature extraction for each segment [2]. Thus, the operation of the automatic

ISBN: 978-1-61804-251-4

text-independent announcer identification includes several stages:

1. Feature extraction.
2. Modeling of speaker.
3. Comparison of the speaker models.

This soundtrack is mapped to the reference speaker soundtrack, by comparing the decision, whether a voice recording belong to this person or different people.

2. FEATURE EXTRACTION

Feature extraction process inherently is not specific to the tasks of speaker identification, but rather is common to most areas of speech technology. For the analysis in the speech signal is assumed to use a set of features such as signal energy, linear prediction coefficients, coefficients of smoothed power spectrum, coefficients of real cepstrum, formant frequencies and pitch period for voiced phonemes. Present correlation between features can be reduced by applying principal component analysis to the vector features.

Energy of signal:

$$E(n) = \sum_{m=-N}^{N} x^2(m) \cdot w(n-m),$$

Where $w(n-m)$ - window function, for example, a Hamming window:

$$h = \left[0.45 - 0.46 \cdot \cos 2\pi n/(N-1)\right], \qquad 0 \le n \le N-1.$$

Linear prediction coefficients, which are the result of solving a system of linear equations Yule - Walker:

$$\sum_{k=1}^{p} a_k R_n(i-k) = R(i), \quad 1 < i <= p,$$

where p – prediction order, $R(i)$ - autocorrelation function, a_k - linear prediction coefficients $1 < k <= p$. Hamming window reduces the prediction error, as the first p samples of a rectangular window with linearly unpredictable.Autocorrelation function calculated with a window:

$$R_n(k) = \frac{1}{N-1-k} \cdot \sum_{m=0}^{N-1-k} \left[x(m)h(n-m)x\right] \cdot \left[x(m+k)h(m-k-m)\right].$$

Formant frequency is determined by the smoothed power spectrum:

$$|H(z)|^2 = \left|\frac{G}{A(z)}\right|^2,$$

where $z = e^{j\omega}$, $H(z)$ - the transfer function of the vocal tract, $A(z)$ - z-transform of linear prediction coefficient sequences.

Cepstral coefficients:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)| \cdot e^{j\frac{2\pi kn}{N}},$$

where $X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi kn}{N}}$ - Fourier transform of signal frame.

Pitch period is determined using the window $l(n)$ for cepstrum:

$$T(n) = c(n) * l(n) \quad l(n) = \begin{cases} 0 \,|\, n| < n_0 \\ 1 \,|\, n| \ge n_0 \end{cases} \quad n_0 = \arg\max(T(n)) \ne 0,$$

where n_0 - pitch period.

Obtained characteristics form the feature vectors $\mathbf{X}$ for each speech segment.

3. MODELING OF SPEAKER

A set of multivariate probability density functions (PDS) describe the hidden acoustic classes of feature vectors. PDS is suitable to approximate arbitrary distributions of the components of acoustic features, making PDS quite convenient for applications in text independent speaker identification and verification.

Usually in the problem of text independent speaker recognition a Gaussian Mixture Models (GMM) are used. GMM is a speaker probabilistic model for multivariate probability density functions (PDS). This model has the obvious disadvantage is that the distribution of acoustic features of speech signals are non-Gaussian, distributions are more peaked. A family of PDS of various shapes, are characterized by three parameters: the expectation m_x, standard deviation σ_x^2 and shape parameter α.

$$f(x) = \frac{\alpha}{2\lambda\boldsymbol{\Sigma}_x \cdot \Gamma\left(\frac{1}{\alpha}\right)} \cdot \exp\left(-\left|\frac{\mathbf{x}-\mathbf{m}_x}{\lambda\boldsymbol{\Sigma}_x}\right|^{\alpha}\right), \quad (1)$$

where $\Gamma(a) \equiv \int_0^{\infty} x^{a-1}\exp(-x)dx$

The distribution function has the form

$$F(x) = \int_{-\infty}^{\left(\frac{x-m_x}{\lambda\sigma_x}\right)^{\alpha}} \frac{1}{2\tilde{A}\left(\frac{1}{\alpha}\right)} \exp^{-\zeta}\zeta^{\frac{1}{\alpha}-1}d\zeta,$$

Scale parameter ß = $\lambda\sigma_x$ of distribution depends on the multiplier λ, which is expressed in

ISBN: 978-1-61804-251-4

terms of the shape parameter α according to the relationship:

$$\lambda = \sqrt{\frac{\tilde{A}\left(\frac{1}{\alpha}\right)}{\tilde{A}\left(\frac{3}{\alpha}\right)}}$$

Centers of GMS are proposed to determine using Radial-Basis Function (RBF) network. Centers of RBF and other parameters of network undergo a supervised learning process. The most convenient for RBF network learning is a gradient descent algorithm that represents a generalization of the Least Mean Square (LMS) algorithm.

The family of RBF networks is broad enough to uniformly approximate any continuous function on a compact set.

Family of RBF networks consists of functions represented by:

$$F(\mathbf{x}) = \sum_{i=1}^{m} a_i \phi(\mathbf{w}_i^T \mathbf{x}) \qquad (2)$$

where m - the number of neurons in the first layer, a_i, $\mathbf{w}_i$ - coefficients of neural network, $\phi(.)$ - the activation function.

As the activation function $\phi(\mathbf{w}_i^T \mathbf{x})$ in the expression (2) a family of exponential distributions (1) with the shape parameter α is proposed.

Calculating the mean square error of approximation of the mixture of multidimensional sampling distributions:

$$\varepsilon = \frac{1}{2}\sum_{j=1}^{N} e_j^2$$

where N – is the size of the training sample .

where N - is the size of the training sample.

Error signal defined by:

$$e_j = d_j - \sum_{i=1}^{M} w_i f(\mathbf{x_j} - \mathbf{m}_i)$$

where d_j – data. The requirement is to find parameters w_i, $\mathbf{m}_i$, $\mathbf{\Sigma}$, α_i.

For better convergence of the algorithm initial values of parameters are selected. Clustering of the sample data is performed according to the method of *k*-means, which estimates initial value of the centers $\mathbf{m}_i$, the initial values of α_i are chosen close to the $\alpha_i = 2$, correlation matrix $\mathbf{\Sigma}$ is chosen close to diagonal, weights are initialized with random values.

Neural network training procedure is performed incrementally. Changing weights on the next step:

$$w_i(n+1) = w_i(n) - \eta_1 \frac{\partial \varepsilon(n)}{\partial w_i(n)} \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial w_i(n)} = \sum_{j=1}^{N} e_j(n) f(\mathbf{x_j} - \mathbf{m}_i(n))$$

Adjustment of the position of the centers:

$$t_i(n+1) = t_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial t_i(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial t_i(n)} = \alpha w_i(n) \sum_{j=1}^{N} e_j(n) f'(\mathbf{x_j} - \mathbf{m}_i(n)) \mathbf{\Sigma}^{-1} (\mathbf{x_j} - \mathbf{t}_i(n))^{\alpha - 1}$$

Adjustment of distribution width:

$$\mathbf{\Sigma}_i^{-1}(n+1) = \mathbf{\Sigma}_i^{-1}(n) - \eta_3 \frac{\partial \varepsilon(n)}{\partial \mathbf{\Sigma}_i^{-1}(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial \mathbf{\Sigma}_i^{-1}(n)} = -\alpha w_i(n) \sum_{j=1}^{N} e_j(n) f'(\mathbf{x_j} - \mathbf{m}_i(n)) \mathbf{Q}_{ij}(n)$$

$$\mathbf{Q}_{ij}(n) = (\mathbf{x_j} - \mathbf{m}_i(n))^{\alpha - 1} (\mathbf{x_j} - \mathbf{m}_i(n))^T$$

Adjustment of the PDS shape parameter:

$$\alpha_i(n+1) = \alpha_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial \alpha_i(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial \alpha_i(n)} = 2 w_i(n) \cdot$$

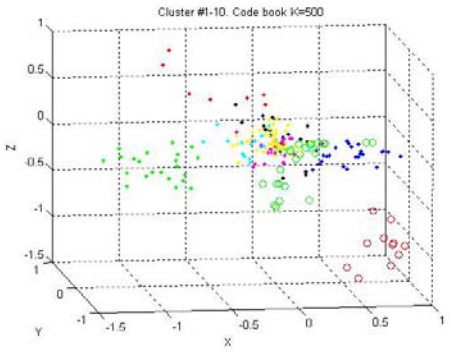
$$\sum_{j=1}^{N} e_j(n) f(\mathbf{x_j} - \mathbf{m}_i(n)) \alpha^{-1} + f'(\mathbf{x_j} - \mathbf{m}_i(n)) \left(\alpha \left| \frac{\mathbf{x} - \mathbf{m}(n)}{\lambda \mathbf{\Sigma}} \right|^{\alpha - 1} \right)$$

4. RESULTS OF EXPERIMENT

The experiment used 15 phonograms recording any text longer than 22000 samples. Analyzed male and female voices same and different speakers, for each phonogram obtained multidimensional histogram features. An example is shown in Figure 1.

ISBN: 978-1-61804-251-4

Figure 1. Projection ща histogram on the three main components of features



K-means obtained initial values of distributions centers. Number of PDS ranged from 10 to 500.

Estimation of errors of the first and second kind for different size of RBF neural network? for different sample sizes and for different speakers, in order to determine the optimal parameters for recording speaker identification.

REFERENCES

[1] A.N. Vasiliev, D.A. Tarkhov Neural Network Modeling: Principles. Algorithms. Applications: Scientific publication/STU. St. Petersburg: Publishing House of STU, 2009, 527 p.

[2] Kotov V.V. Automatic text speaker identification based on a telephone conversation. Science Week XXXIX STU: Proceedings of the International Scientific and Practical Conference. Charles VIII. - St. Petersburg. Univ Polytechnic. University Press, 2010, p. 122-124.

[3] Malykhina G.F. Engineering and technical protection of information: Speech Technology: Textbooks / STU. St. Petersburg: Publishing House of STU, 2004, 243 p.

[4] Pervushin E.A. Basic methods for speaker recognition / / Mathematical Structures and Modeling. - Omsk, 2011, NVyp. 24. - S. 41-54.

[5] Sorokin V.N., V'yugin V.V., Tankin A.A. Information technology in the technical and socio-economic systems. Individual voice recognition: analytical review / / Information Processes. - Moscow, 2012, Volume 12, N 1. - p. 1-30

ISBN: 978-1-61804-251-4