Speaker Identification Using Autoregressive Hidden Markov Models and Adaptive Vector Quantization

Eugeny E. Bovbel, Igor E.Kheidorov, Michael E.Kotlyar

Dept. of Radiophysics, Belarussian State University
F. Scoriny, 4, 220050, Minsk, Belarus
tel, fax. (0172) 770-890
e-mail: ikheidorov@poboxes.com

20 May 2000

Abstract

Wide-frequency spectral analysis, autoregressive hidden Markov models (ARHMM) and self-organizing neural networks (SOM) have been used for high accuracy speaker features modelling. The initial ARHMM parameters estimation based on Kalman filter is proposed. The five-keyword speaker identification system has been built and tested. The experiments show that this approach provides high accuracy of speaker identification even if the same words are pronounced by different speakers.

1 Introduction

By its nature speech signal has two aspects. Firstly, it is highly dependent on physical parameters of vocal tract. Secondly, the speech is produced under neural control of the human, so that it is affected by the person-specific characteristics such as accent, pronunciation, speed, timbre etc. It gives us the possibility to use these features for speaker identification.

We model the person-specific characteristics taking into consideration the speech signal parameters interrelation between close frames. This approach provides us with the useful information about pronunciation mode. From this viewpoint, statistical methods based on autoregressive HMMs are very attractive, because the standard HMM does not care about temporary close observations interrelation, which makes valued information about acoustical structure of the phoneme loose.

2 Autoregressive hidden Markov models

The speech utterance can be modelled by the sequence of the discrete stationary states with immediate transitions between them. Let $\overline{O} = (O_1, O_2, ..., O_N)$ be the observation vector sequence for the given utterance, and $\overline{q} = (q_1, q_2, ..., q_N)$ be the appropriate HMM states sequence. $O_n = (x_{n,1}, x_{n,2}, ..., x_{n,K})$ - nth K-dimensional observation vector. For each state q it is necessary to define distribution of observation vectors $b_q(O)$. We assume that the vector sequence $\overline{O}$ components conform the P-order autoregessive model:

$$x_{n,k} = -\sum_{i=1}^{P} a_{k,i} x_{n-i,k} + \epsilon_k \tag{1}$$

where ϵ_k - independent zero-mean Gaussian random values with dispersion σ^2; $a_{k,i}$ - autoregression coefficients (linear prediction coefficents) for kth component of observation vector.

With this assumption for each state q we can estimate observation vector components distribution as following:

$$\begin{aligned} & b_q(x_{n,k} | x_{n-1,k}, x_{n-2,k}, ..., x_{n-P,k}) = \\ = & \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{1}{2\sigma_q^2}(x_{n,k} + \sum_{i=1}^{P} a_{k,i,q} x_{n-i,k})^2\right) \end{aligned} \tag{2}$$

So, the autoregressive hidden Markov model is a model of a twice stochastic process: the model state sequence is first-order Markov process, and the observation vectors sequence is random process too and is modelled by P-order autoregression.

3 Initial estimation of ARHMM parameters

We have used Kalman filter to estimate initial autoregression parameters for ARHMM. Consider linear system. States of the system in moment t and in moment $t-1$ are constrained by the system equation:

$$q_t = F q_{t-1} + N_t \tag{3}$$

where F - linear operator, N - system noise.

Let q_t - linear system state in moment t. At every moment t we perform measuring of system output y_t, wich is function of system state:

$$y_t = H q_t + G_t \tag{4}$$

where H - measuring operator, G_t - measuring noise. If we suppose $q_t = (a_1, a_2, ..., a_P)$, $y_t = x_{t,k}$, we can rewrite 3,4 as following:

$$q_t = q_{t-1} + N_t \tag{5}$$

$$y_t = H_t q_t + G_t \tag{6}$$

$$H_t = -diag(y_{t-1}, y_{t-2}, ..., y_{t-P}) \tag{7}$$

Using Kalman algorithm, if y_t sequence is given, we can estimate system parameters q. The detailed description of Kalman filter can be found in [6].

4 Experiment

Five belarussian words spoken by 5 speakers have been sampled at 44kHz rate in different environmental conditions during 3 months. 130 samples per word per dictor have been taken, 3250 samples totally. Samples have been separated in two nonintersecting sets: the training and the testing one. Mel-cepstrum and delta-cepstrum have been used as observation vectors. Separated SOM-codebooks and mixed discrete-continuous ARHMMs have been built for cepstrums and delta-cepstrums. Probabilities given by corresponding cepstr- and delta-cepstr ARHMMS were multiplied. Two codebooks (for cepstrums and delta-cepstrums) were trained on all vectors made from training set. Then, ARHMM pairs have been trained on the training set - one ARHMM pair per word per dictor, 25 model pairs totally.

First series of experiments have been carried out to determint the optimal codebook size (see Fig.1).

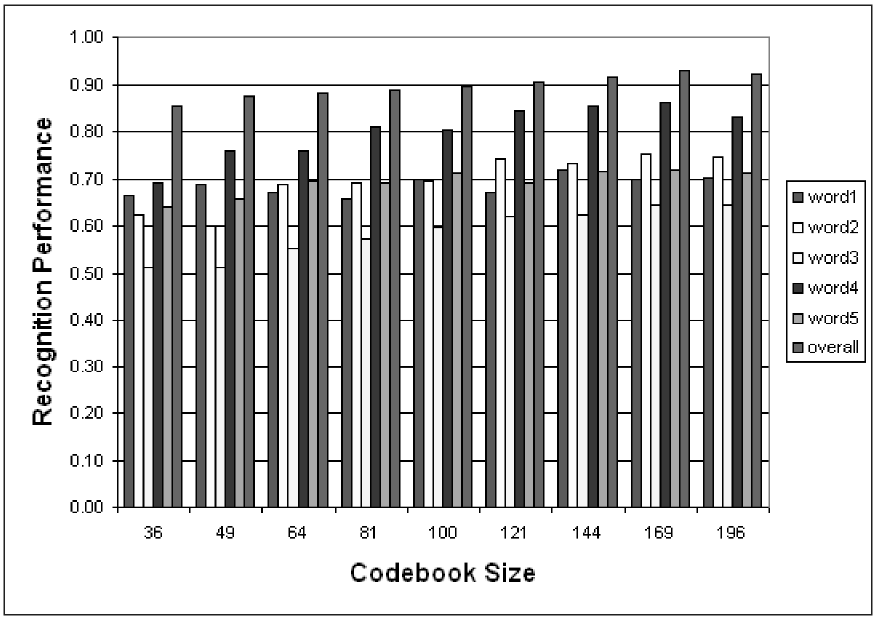

*Fig.*1

Training set consisted of 5 samples per word per speaker, remaining samples were used for testing. As we can notice, optimal codebook size is about 139, which is rather much as usual in word-recognition tasks.

Second series of experiments was designed to estimate our speaker recognition system's overall performance (see Fig.2).

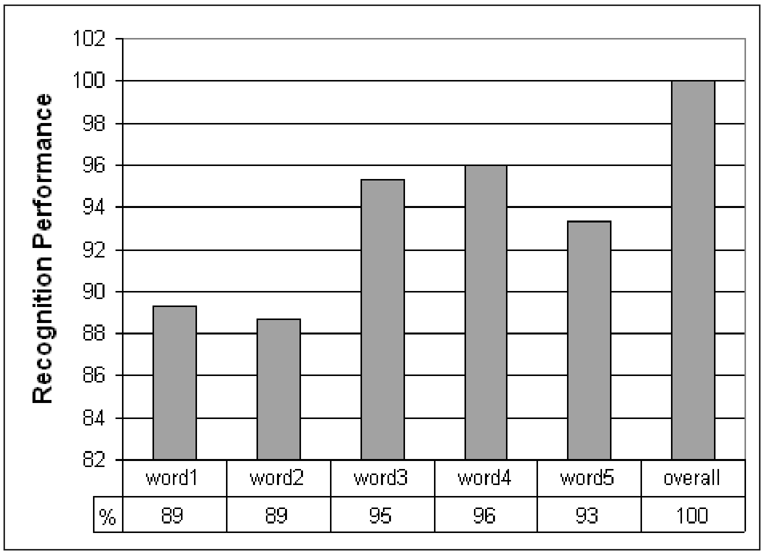

*Fig.*2

Training set consisted of 30 samples per word per speaker, remaining samples were used for testing. As we can see, overall recodnition performance, when every speaker pronounces the sequence of the same 5 words, is close to 100%.

References

[1] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Comunication, 17(1-2), pp.91-108, 1995

[2] H. A. Bourlard and N. Morgan Connectionist Speech recognition: A Hybrid Approach. Kluwer Academic Publishers, Boston M.A., 1994

[3] B.H. Juang, L.R.Rabiner, Mixture autoregressive hidden Markov models for speech signals, IEEE Trans. ASSP-33, 6, pp.1404-1412, 1985

[4] E.I.Bovbel, I.E.Kheidorov, P.P.Tkachova, The analysis of speaker individual features based on autoregressive hidden Markov models, Proc. of EUROSPEECH'99, vol.3, pp.1191-1194, September 5-9, 1999, Budapest, Hungary.

[5] T.Kohonen, The self-organizing map, Proceedings of IEEE, vol.78, 1990. pp. 1464 - 1480.

[6] Youji Iiguni, A real-time learning algorithm for a multilayered neural network based on the extended Kalman filter, IEEE Transactions on Signal Processing, vol.40, April 1992, pp.959-966.