Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

AUTOREGRESSIVE HIDDEN MARKOV MODEL FOR APPLIED TASKS OF VOCAL FOLD PATHOLOGY DETECTION.

Evgeny I.Bovbel, Polina P.Tkachova, Igor E.Kheidorov
Dept. of Radiophysics, Belarusian State University,
F.Scoriny av., 4, 220050, Minsk, Belarus
e-mail: ikheidorov@poboxes.com

Abstract. The recognition and training algorithms for autoregressive hidden Markov models were developed in order to solve the task of vocal fold pathology detection. Three databases were created and used for 3 vocal pathologies detection. During the experiments the proposed vocal tract pathology detection system based on autoregressive hidden Markov models and wide-range AFT mel-spectrum provides very high detection accuracy.

INTRODUCTION

The analysis of speaker individuality is widely used for different tasks, for example, speaker identification and verification, diagnostics of speech producing organs, secure access [1], [2]. It is well known that if there is the presence of vocal fold pathology, significant changes can occur that alter the speech production system, resulting in deterioration of voice quality. Analysing such changes we can make a conclusion about the state of person's vocal fold and compare it with the template health person parameters. Subjective testing made by a physician is not able to detect pathology on the earlier stages, except this such testing strictly depends on the physician experience. There is a possibility to perform such analysis objectively based on more or less nonlinear model. From this point of view systems based on hidden Markov models (HMM) are very attractive. Flexible and powerful mathematical apparatus of these models lets to use them for effective temporal information modelling. Speech signal by its nature has two aspects. Firstly, speech of the person is defined by physical parameters, such as vocal tract length, glottal size and so on. Secondly, the speech producing is impossible without neural control of the articulators, which defines the personal learned abilities such as dialect or regional accents, pronunciation, speed and timing of the articulators.

For better analysis of speaker pronunciation peculiarities it is necessary to take into consideration interrelations between close frames of the same phonetic unit and loose important information about acoustical structure of the phoneme. In connection with this we introduce an autoregressive hidden Markov model for the task of vocal tract pathology detection, which is similar to the task speaker identification within the framework of such approach [3]. The character vector used for speech analysis greatly influences the voice analysis performance. In order to provide the high accuracy we have to use apriory knowledge about human speech producing and perception. In particular, the knowledge of main psycho-acoustical principles gives us the possibility to cut information not grasping by human ear. Information about high frequencies in the speech signal is very important also, and it is rather reasonable to use it for vocal fold pathology detection.

SPEECH PARAMETERS FOR VOCAL FOLD PATHOLOGY DETECTION

Now human ear is the best tool for speech analysis and it is not a bad idea to model it in order to solve different speech analysis tasks. The usage of psycho-acoustical principles lets to form speech parameters that reflect all essential speech features. Psycho-acoustical parameters include absolute threshold of hearing, critical bands, simultaneous and temporal masking [4]. Here we briefly review the psycho-acoustical parameters in order to include them to the speech feature vector estimation process.

The absolute threshold of hearing is characterized by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The threshold is well approximated by the nonlinear function

$$T_q(f)=3.64(f/1000)^{-0.8}-6.5e^{-0.6(f/1000-3.3)^2}+10^{-3}(f/1000)^4 \quad (1)$$

Using the absolute threshold of hearing represents the first step towards perceptual speech analysis. For the next step it is useful to consider how the ear actually does spectral analysis. It turns out that a frequency-to-place transformation takes place in the inner ear, along the basilar membrane. Different regions in the cochlea are "turned" to different frequency bands. This process was deeply investigated and the term "critical band" was established. In the experimental sense, critical bandwidth can be loosely defined as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise source at constant sound pressure level remains constant even as the bandwidth is increased up to the critical bandwidth. For an average listener, critical bandwidth is conveniently approximated by

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69}(Hz) \quad (2)$$

Although the function BW_C is continuous, for practice it is useful to present the ear as a discrete set of bandpass filters. A distance of 1 critical band is referred to as "one bark". In order to convert frequencies from Hertz to te bark scale we use the function

$$z(f) = 13\arctan(.00076f) + 35\arctan\left[\left(\frac{f}{7500}\right)^2\right](Bark) \quad (3)$$

Often for the discrete set of filters the following approximation of this function called mel-scale, is used

$$Mel = 2595 \log_{10}(1 + f/700) \quad (4)$$

Such psycho-acoustical principles can be used for all speech analysis tasks, for example, for speech recognition. It was shown that for the better personal features modelling it is necessary to use high frequency domain (above 10 kHz) and high dynamical range for the representation of input speech signal (not less than 16 bits/sample). The effectiveness of such approach is obvious because for high-accuracy speech analysis tasks we can use necessary information about high frequencies without significant processing time increasing.

To provide the high accuracy of signal processing and wide dynamical range it is reasonable to use spectrum estimation methods with low level of calculation errors. One of such methods is based on the arithmetic Fourier transform (AFT) [5], [6]. This choice can be explained by several reasons. The main drawbacks of using FFT for speech recognition are that it is impossible to use free analysis window length and there is the strict relationship between the speech sampling frequency and the FFT spectral resolution. AFT has no restrictions for the number of input points and do not require complex component calculations. That is why there is no the necessity to store even/odd coefficients and use complex memory addressing typical for the most FFT algorithms. The AFT properties make it to be very attractive for the VLSI realization to solve the tasks of signal Fourier analysis, in particular, speech recognition tasks. Especially we want to note that AFT is very suitable from the viewpoint of perceptual speech signal processing. Mel-scale frequencies and thresholds can be easily defined using AFT, much more convenient than using FFT.

The accumulated AFT-based spectrum values within the certain frequency range (channel) can be used as speech parameters:

$$S_{avg} = \frac{1}{N}\sum_{n=0}^{N} \omega(f_{mel}, n) S(f_{mel}), \quad (5)$$

where N - number of spectral samples, $\omega(f_{mel}, n)$ - weight function.

Let S_i - spectral power of within the channel i. Then $\{S_1, S_2, \ldots, S_L\}$- vector of speech parameters, where L- number of used channels. AFT spectrum can be estimated with any accuracy within the necessary range, therefore this method of spectrum estimation is very attractable to form mel-parameters. To increase the robustness we can perform the cos transformation of speech parameters. This transformation forms coefficients like the cepstral coefficients, but they are based on other principals. Thus, using the values of signal energy within the spectrum range $E_i = log(S_i)^2$ we obtain the "mel-cepstrum":

$$C_k = \sum_{i=1}^{L} E_i\, cos(k(i - 1/2)\pi/L), \; 1 \le k \le K, \quad (6)$$

where K- desired number of "mel-cepstrum" values.

HIDDEN MARKOV MODEL BASED ON AUTOREGRESSIVE PRINCIPLES.

Speech signal is produces by a non-stationary nonlinear process, and for analysis we need to present it as part-stationary signal. In this case we can apply a statistical model for speech features within the quasi-stationary parts of the signal. The main task of hidden Markov modelling is to estimate the probability that the given speech signal (process of feature vectors observation) is generated by the given hidden Markov model (process of transitions between hidden stationary states with certain statistics).

Let we have a certain word spoken by M speakers constructed from L phoneme-like units. Let $Q^* = \{1,2,...,l,...,L\}$ is a set of HMM states, which are used to build word models. The word for given speaker can be modelled by the sequence of L_i discrete stationary states $q_n \in Q^*$, $L_i \le L$ with immediate transitions between states. For each word HMM defines:

- stationary state sequence $Q = \{q_1, q_2, ..., q_n, ..., q_N\}$ which models the temporal speech structure (hidden process);
- acoustic vector sequence $O = \{O_1, O_2, ..., O_n, ..., O_N\}$ which models local-stationary properties of speech signal (observable process).

Formally the hidden Markov model for the speaker m, $m = 1,2,...,M$, can be indexed as $\lambda_m = (\pi, W, B)$, where

1) $W = (w_{ij}) = P(q_n = j \mid q_{n-1} = i)$- state transition probabilities matrix with size $L_m x L_m$,

2) $B = \{b_j(O_n)\}$- probability distribution of acoustic observation vector appearance at state j.

3) $\Pi = (\pi_i) = P(q_1 = i)$- initial probabilities matrix with size $L_m x1$.

To design a set of hidden Markov models for person identification we assume:

- the given word model is presented by the sequence of elementary speech units (phones and phonemes). Discrete stationary HMM states are directly connected to the elementary speech units.

- the word of each person is presented by a single hidden Markov model, which described the statistical variations in word pronunciation speed.

- acoustical signal belonged to the given speaker is a realization of part-stationary process (within 20-70ms intervals speech signal is assumed to be a stationary one).

To use this technique for speech recognition we have to develop algorithm for the estimation of aposteriory probability $P(\lambda_m \mid O)$, $m = 1,2,...,M$. The speaker w which provides the maximum $P(\lambda_m \mid O)$ is to be accepted as the identification result:

$$w = \underset{m \in M}{\operatorname{argmax}} P(\lambda_m \mid O). \qquad (7)$$

In general case we can only estimate the probability $P(\lambda_m \mid O)$ using given data and HMM assumptions. For this purpose we apply Bayesian rule which presented $P(\lambda_m \mid O)$ as

$$P(\lambda_m \mid O) = \frac{P(O \mid \lambda_m) P(\lambda_m)}{P(O)}. \qquad (8)$$

The calculation of aposteriory probability has two stages: probability likelihood function $P(O \mid \lambda_m)$ calculation for the given model λ_m which depends on the acoustical data, and apriory probability $P(\lambda_m)$ estimation. At the identification stage $P(O)$ is constant as the rule, $P(\lambda_m)$ can be easily found based on the linguistic language analysis, that is why the main task is to calculate $P(O \mid \lambda_m)$, which is called the acoustical likelihood function. We have to provide effective and fast calculation algorithm for $P(O \mid \lambda_m)$ using the given series of acoustical vectors $O = \{o_1, o_2, ..., o_N\}$ and hidden Markov model λ_m.

Let discuss the conditions to solve the speaker identification task based on hidden Markov model and autoregressive principles [3]. The main task of identification is to estimate the acoustical probability likelihood function $P(O_1 O_2 ... O_N \mid \lambda_m)$ for different HMMs $\lambda_m, m = 1,2,..., M$. Let perform the calculation of acoustical probability likelihood function using the popular forward-backward procedure. Then we suppose that the assumption $P(O_n, q_n \mid q_{n-1}, ..., q_1, O_{n-1}, ..., O_1) = P(O_n, q_n \mid q_{n-1})$ typical for standard HMM is not valid more. The probability of acoustical vector O_n appearance depends on not only the current state $q_n = j$, but the previous states $q_{n-1} = i, q_{n-2}, ..q_{n-1}$ too. This assumption suits the real speech signals. Such dependence can be linear and nonlinear. In order to decrease the number of free parameters for HMM we use only linear dependence.

Let's consider the state sequence $Q = \{q_1, q_2, ..., q_n, ..., q_N\}$ and appropriate observation vector sequence $O = \{O_1, O_2, ..., O_n ..., O_N\}$, where $O_n = \{x_1, x_2 ... x_k, ... x_K\}$- observation vector, which consists from K parameters. For each state $q_n = j$ it is necessary to define $b_j(O_n)$. We assume that the vector sequence O components suit to the p-order autoregressive model $AR(p)$:

$$x_n = -\sum_{i=1}^{P} a_i x_{n-i} + \varepsilon_n , \tag{9}$$

where ε_n-Gaussian independent random values with zero-mean and dispersion σ^2, a_k-autoregressive coefficients (linear prediction coefficients).

Using this assumption for each state $q_n = j$ at time n we can write the following expression

$$b_j\left(x_n \mid x_{n-1}, x_{n-2}, \ldots, x_{n-p}\right) = \frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma_j^2}\left(x_n + \sum_{i=1}^{P} a_i^j x_{n-i}\right)^2\right\} \tag{10}$$

It is possible to show that for large K the probability density $b_j(O_n)$ can be written in following manner

$$b_j(O_n) = \prod_{n=1}^{K} b_j\left(x_n \mid x_{n-1}, \ldots, x_{n-p}\right) = \frac{1}{\left(2\pi\sigma_j^2\right)^{k/2}} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{n=1}^{K}\left(x_n + \sum_{i=1}^{P} a_i^j x_{n-i}\right)^2\right\} \tag{11}$$

In such a way the autoregressive hidden Markov model is a model of twice stochastic random process. Firstly, the model state sequence which presents the temporal speech structure, is the sample function of first-order Markov process. Secondly, the observable process which is defined by local properties of speech signal, is the random process too and it is modelled by p-order autoregression process $AR(p)$.

EXPERIMENTS AND CONCLUSION

For each pathology we train two hidden Markov models using speech samples produced by ill people and health people. To detect the pathology we have to estimate the aposteriory probability $P(\lambda_m \mid O)$, $m = 1,2$, where λ_m- hidden Markov model, $O = \{O_1, O_2, \ldots, O_n, \ldots, O_N\}$- acoustic vector sequence, for each model. The model ξ which provides the maximum of $P(\lambda_m \mid O)$ is to be accepted as the analysis result according to (7).

The recognition and training algorithm for autoregressive hidden Markov models were developed in order to solve the task of vocal fold pathology detection. Three databases were created and used for 3 vocal pathologies detection. 5 phones were spoken by health and ill speakers 20 times and sampled at 44kHz. Speech character vector formed from 24 spectrum coefficients in mel-scale and covered range from 100Hz upto 20kHz, was used for the experiments. Arithmetic Fourier transform (AFT) was applied as spectrum estimation method because of high accuracy and suitability for mel-scale implementation [4], [5]. Experiments show that high frequencies of speech signal spectrum are very sensitive to speaker individuality, and that is why we use so wide frequency range. Autoregressive hidden Markov models were trained for each of three tested pathologies. During the experiments the proposed vocal tract pathology detection system based on autoregressive hidden Markov models and wide-range AFT mel-spectrum provides very high detection accuracy.

REFERENCES

1. H.Bourlard, N. Morgan Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers,-1994.

2. D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models //Speech Communication, №17(1-2), pp.91-108, 1995

3. B.H. Juang, L.R.Rabiner, Mixture autoregressive hidden Markov models for speech signals //IEEE Trans. on Acoustics Speech and Signal Processing, ASSP-33, №6, pp.1404-1412, 1985

4. E.I.Bovbel, I.E.Kheidorov, Speech Parameters Vector Based on Arithmetic Fourier Transform //Proc. of XI European Signal Processing Conference EUSIPCO 98. pp.713-717, 1998

5. V.G.Atlas, D.G.Atlas, E.I.Bovbel, 2-D arithmetic Fourier transform using Bruns method //IEEE Trans. on Circuits and Systems, vol.44, №6, pp.546-552, 1997.

This paper is partialy supported by the International Science and Technology Center, project B-95.