Stochastic speaker recognition model

Július Zimmermann and Július Zimmermann, Jr.

The task of speech processing is being spread over last years with speaker recognition, which consists of speaker identification and speaker verification. In the paper there is an explained method which uses the stochastic model to solve this task. The experiment has been done using hidden Markov models by means of HTK tool software. Testing speech materials consist of 200 speakers from corpus SpeechDat-Sk. Applied experiments reached the efficiency between 92,86% and 100%.

1. Introduction

Human voice timbre and quality belong to unique speaker qualities. Already in the past they have served for distinguishing different people. Acoustic feature differences of speech signal of an individual are results of an anatomic difference combination in the articulatory tract and learnt individual speaking habits. In present rapid development and during putting automated systems into existence they may be applied also in the devices for speaker recognition, for instance authorised check in, banking via telephone, and in the telephone credit cards.

Speaker recognition involves verification and identification. At verification the voiceprint is compared with the speaker voice model registered in the database that we want to verificate. The result of comparison is a measure of the similarity (score) from which rejection or acceptance of the verified speaker follows. At identification the voiceprint is compared with model voices of all speakers in the database. The comparison results are measures of the similarity from which the maximal quality is chosen.

Following factors influence errors at speaker recognition: an incorrect reading or incorrect expressing of the required phrase, extreme emotional state, the changed microphone position within individual recognition or between individual recognitions, a different microphone at the speaker enrolment into the database and at the recognition, room acoustics, disease of the vocal tract and age. The speaker recognition system can be put together in a way to be resistant to noise, voice changes and imitation of the speaker; on the other hand, it is easier to guarantee the constant acoustic conditions at the microphone than to increase the robustness of the system.

There are several methods for speaker recognition which classify coordination (or dissimilarity) of the voiceprint and speech model. They can be divided into template methods, stochastic methods and methods which use artificial neural networks. Into template methods we rank Dynamic Time Warping, Vector Quantization and the combination of both mentioned methods, Nearest Neighbours. Hidden Markov Model belongs to stochastic models and it is also used in the experiment that this paper refers to.

The speaker recognition consists of these phases: digital speech data acquisition, feature extraction, creation and speaker model recording into the database (training), matching model with voiceprint (testing) and the test classification (Figure 1).

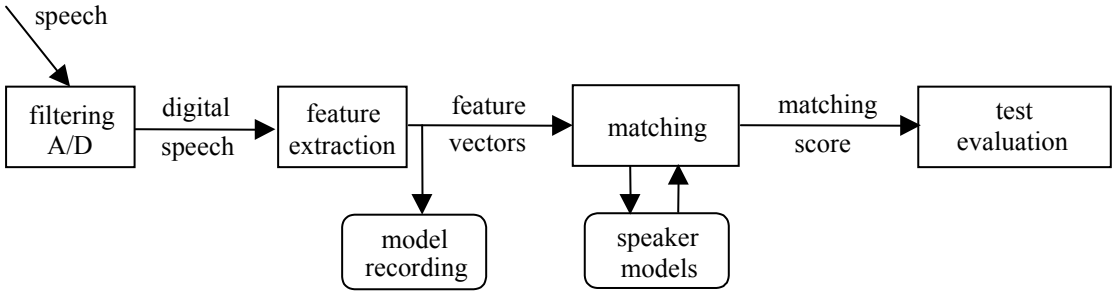

Figure 1

2. Mel-frequency cepstral coefficient extraction

In this phase the digital speech signal is partitioning into segments (frames) with fixed length 10-30 ms from which the features are extracted due to their spectral qualities. Spectrum is achieved with fast Fourier transformation. Then an arrangement of frequency range to mel scale follows according to relation (1):

$$f_{mel} = 2595.\log\left(1+\frac{f_{Hz}}{700}\right) \tag{1}$$

By logarithm of amplitude of mel spectrum and applying reverse Fourier transformation we achieve frame cepstrum:

$$mel-cepstrum(frame) = FFT^{-1}[mel(\log|FFT(frame)|)] \tag{2}$$

The features calculated for each frame are usually expanded with their dynamic qualities by first and second features derivation.

3. Stochastic speaker models

Stochastic models show flexibility and probability. In this method the particular model is dedicated to every speaker, the parameters of which are determined by its training data. When an unknown speaker utters some phrase which is then parameterised on the feature vectors $O = o(1), o(2), ...$, so the speaker identification i is determined by the relation:

$$i' = \arg\max_{i \in I} P(i \mid O) \tag{3}$$

where *i'* is the identified speaker from the speakers *i* in the whole speaker set *I*. It means that the probability that the generated vector *O* belongs just to the speaker *i* is firstly determined. Then the speaker is identified on the basis of this probability in a way that the speaker with maximal probability is chosen. Direct counting of this probability is very complicated. The calculation is simplified by breaking the probability down to partial probabilities according to Bayes as follows:

$$P(i \mid O) = \frac{P(O \mid i)P(i)}{P(O)} \tag{4}$$

it is assumed that the probability $P(i)$ of the speaker utterance *i* is known (usually it is the same for all speakers). $P(0i)$ expresses the probability of generating the feature *0* by the speaker *i* and $P(0)$ is the probability of generating the feature *0* by any speaker. This probability is the mean value of probability $P(0|i)P(i)$. Above mentioned correlation can be understood as comparison of the probability generating the feature by the speaker *i* to average probability of generating of the observed feature by all speakers. After simplification it is enough to measure probability $P(0|i)$ of observation the sequence of feature vectors *0* from unknown speaker *i*. Stochastic speaker models serve to it, and by them we can determine the probability.

3.1. *Hidden Markov Models – HMM*

HMM are very popular for sequence modelling. The Figure 2 HMM represents five states. HMM are finite state models and the system can occur in only one of these states at any given moment. HMM represents two stochastic processes. The transition process between the states and stochastic process of generating feature vectors by individual states. The first mentioned stochastic process of sequence states is not directly observable (it is hidden), just the process of generated feature vectors is observed. The states are linked with the passage network where the probability of the transition from the state *i* to the state *j* are a_{ij}. In each time leap between the speech frames the model changes its state from *i* to *j* with the probability a_{ij} and generates from the state *j* the feature vector $o(i)$ with the probability $b_j(o(i))$. All states are emitting with the exception of the first and last one, which stand for connecting the models.

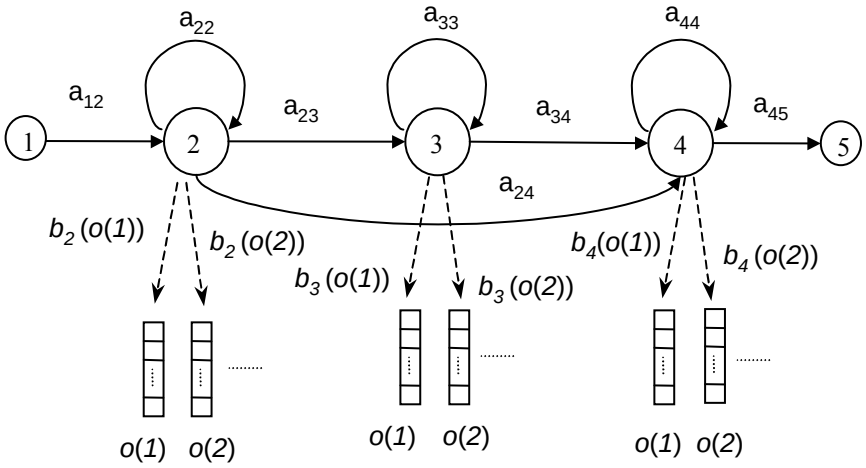

Figure 2

HMM is characterised by these parameters:

1. States number in a model – N.
2. Feature vectors number generated by one state – M.
3. Matrix of the probability passage between the state $\boldsymbol{A}=\{a_{ij}\}$ where $1 \le i, j \le N$. If a passage from any state to any other state by one step exists, then it is ergodic HMM and $a_{ij} > 0$ for all i, j is valid. This condition does not stand for other types HMM.
4. Matrix $\boldsymbol{B}$ of probabilities $b_j(o(i))$, $\boldsymbol{B}=\{ b_j(o(i))\}$ where $1 \le j \le N$ and $1 \le i \le M$.
5. Vector of initiation probability by i-state $\pi_i = P(q_1 = s_i)$ for $1 \le i \le N$, it expresses the probability that the initial state will be the state i.

The probability $b_j(o(i))$ may be defined by discrete or continuous probability density function.

Assuming that entry phrase is parameterised into feature vectors $O=o(1)$, $o(2)$, $...o(T)$ and state sequence is X = 1, 2, 2, 3, 3, 4, 4, 5, then the probability of vector generating 0 by the model M through the state sequence X is given as follows:

$$(5) \qquad P(O, X \mid M) = a_{12} b_2(o(1)) . a_{22} b_2(o(2)) . a_{23} b_3(o(3)) . a_{33} b(o(4)) ...$$

As the state progress is unknown in practice (hidden HMM), there are two methods of probability calculation. The first way is that the probability amount of vector generating O by model M through all the possible state sequences are figured out:

$$(6) \qquad P(O \mid M) = \sum_{\substack{all\ state \\ sequences}} a_{x(0)x(1)} \prod_{i=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

The result is Baum-Welch probability.

The other probability calculation uses Viterbi algorithm for finding the most probable states sequence; taking this sequence into account Viterbi probability is determined:

$$\hat{P}(O \mid M) = \max\left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\} \tag{7}$$

Because the model represents the speaker, we may determine the necessary probability $P(O|i)$ for the probability calculation $P(i|O)$ provided:

$$P(O \mid i) = P(O \mid M_i) \tag{8}$$

4. Experiment

The aim of the experiment was to verify the usage possibilities of the speech corpus SpeechDat-Sk at speaker recognition by help of program environment Hidden Markov Toolkit – HTK. Above-mentioned corpus was created with the assistance of Inco-Copernicus project, which is sponsored by the European Committee. The corpus records were made in the Slovak Academy of Sciences in Bratislava who is the 50% legal owner. The Lernout&Houspie Speech Products Company owns other 50%. The corpus contains records of 1000 speakers, which were recorded directly through public telephone network by the use of digital ISDN. The corpus is distributed on 5 CDs. In this experiment there was the first CD used with records of the first 200 speakers.

The speakers were reading numerals, numbers, dates, time, Christian names, and town names during the recording. They were spelling letters, reading 12 phonetically rich sentences. Sampling frequency 8 kHz at the sampling size 8 bites (standard G.711) according to the A-law digitalized the acoustic speech.

Program environment Hidden Markov Toolkit – HTK used was developed by Steve Young at Cambridge University in 1989. The program became the property of the research laboratories of Entropic Company, later Microsoft Company. The used version 3.2.1 from the year 2003 is determined for the work with hidden Markov models in the speech recognition applications. In our experiment this program extracted features, automatically determined the borders of phones, or triphones, created speaker models and compared models in the test processes.

HTK represents the set of executable files, which are written in the language C and compiled on *.exe files. Individual files are executed from DOS prompt with appropriate parameters. Each *.exe file performs some operation. Input for *.exe is – besides the other – also the text file, which contains for instance topology of HMM model or the list of monophones. During the system activity it is necessary to make some changes in the text files, e.g. to form new files and so on; for such a task we used the programs in the Perl language, so called Perl scripts. The Perl language is extremely suitable for such tasks because it has strong functions for text manipulation of files and abilities to perform boundary between operating system and the user. By scripts use we automated the system work for speaker identification.

5. Results

Experiment results are presented in Table 1:

Experiment number	Amount of speakers	Training	Test	Segment	Success
1	196	complete rec.	5 numerals	triphones phones	99,49 % 99,49 %
2	196	complete rec.	1 numeral	triphones phones	98,97 % 97,96 %
3	196	complete rec. besides test numeral	1 numeral	triphones phones	92,86 % 94,90 %
4	195	all numeral rec.	5 numerals	triphones phones	100,00 % 100,00 %
5	195	all numeral rec.	1 numeral	triphones phones	100,00 % 100,00 %
6	195	all numeral rec. besides test numeral	1 numeral	triphones phones	95,38 % 95,90 %

Table 1

The corpus SpeechDat-Sk is dedicated to speech recognition, so it can be used only for applications of text dependent speaker recognition. If a new corpus demanding the highest successful speaker recognition should be developed it is suggested to enable shaping (model) entire words (or phrases). The shaping (modelling) entire word (phrases) is more precise, and detailed. If we want to enlarge grammar testing with new words the training of new models is necessary that seems to be the disadvantage. In the systems focused on the smaller units training new words are compiled from the already trained units. Additional training of further speech units is not necessary. Flexibility of the system is reached at the expense of precision. In the systems of shaping entire words the speaker should record the same words (or phrases) several times. In addition, from the phonetic aspect those words should characterise speaker features such as formant, basic tone, etc.

Július Zimmermann
Phonetic Laboratory
Faculty of Arts
Prešov University
Slovakia

zimmer@unipo.sk

Július Zimmermann, Jr.
Phonetic Laboratory
Faculty of Arts
Prešov University
Slovakia

zimmermann@centrum.sk

References

Campbell, Joseph P. Jr. 1997. *Speaker Recognition: A Tutorial.* In: *Proceedings of the IEEE*, Vol. 85/9, 1437-1462.

Černocký, J. 1998. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*. Paris, De L´ Université Paris XI Orsay. PhD Dissertation.

Gan, Seng Ern. 2000. *User Authentication By Voice Final Draft*. Nedlands, The University of Western Australia, Faculty of Engineering and Mathematical Sciences. CIIPS Honours Dissertation.

Hemant, Misra. 1999. *Development of Mapping Feature for Speaker Recognition*. Madras, Indian Institute of Technology, Department of Electrical Engineering. MA thesis.

Olsen, Jesper Ostergaard. 1997. *Phoneme Based Speaker Recognition*. Aalborg, Aalborg University, Center for Person Communication. PhD Dissertation.

Magimai-Doss, Mathew. 1999. *Combining Evidence from Different Classifiers for Text-Dependent Speaker Verification*. Madras, Indian Institute of Technology, Department of Computer Science and Engineering 1999. MA thesis.

Pellom, Bryan L. 1998. *Enhancement, Segmentation, and Synthesis of Speech With Aplication to Robust Speaker Recognition*. Duke University, Department of Electrical and Computer Engineering. PhD Dissertation.

Psutka, JOSEF. 1995. "Komunikace s počítačem mluvenou řečí." Praha: Academia, 127-200.

Van Vuuren, Sarel. 1999. *Speaker Verification in a Time-Feature Space*. Oregon, Oregon Graduate Institute of Science and Technology. PhD Dissertation.

SpeechDat: Databases for the Creation of Voice Driven Teleservices. <www.SpeechDat.org>

Sarma, Sridevi Vedula. 1997. *A Segment-Based Speaker Verification System Using SUMMIT*. Massachusetts, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science. MA Thesis.

Taischi, Chi. 1997. *Aplication of Auditory Representations on Speaker Identification*. Maryland, University of Maryland at College Park, Faculty of the Graduate School. MA Thesis.

Kao, Y.-H. 1992. *Robustness Study of Free-Text Speaker Identification and Verification*. Maryland: The University of Mariland, Faculty of the Graduate School 1992. PhD Dissertation.

Young, Steve; Odell, Julian; Ollason, Dave et al. 2002 *The HTK Book (for HTK Version 3.2)*. Cambridge University Engineering Department. <htk.eng.cam.ac.uk>

Zimmermann, J., Jr. 2004. *Automatické rozpoznávanie hovoriaceho pomocou hlasového odtlačku*. Košice, Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach 2004. MA Thesis.