SPEAKER AGE RECOGNITION BASED ON ISOLATED WORDS BY USING SVM

Mengdi Yue[1], Ling Chen[2], Jie Zhang[2], Hong Liu[3]

[1]School of Electronic and Information Engineering, University of Science and Technology LiaoNing, CHINA.
[2]Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School of Peking University, CHINA.
[3]Engineering Lab on Intelligent Perception for Internet of Things (ELIP), The Key Laboratory of Machine Perception, Peking University, CHINA.
yuemengdi521@163.com, chenling@sz.pku.edu.cn, zhangjie827@sz.pku.edu.cn, hongliu@pku.edu.cn

Abstract: Speaker age recognition is an essential technique in automation speech recognition based on the speech wavform parameters in speaker's voice. However, there are several challenges in speaker age recognition, such as innate differences in speaker's voice, subjective classification fuzzy, etc. The issue of speaker age based on isolated words is proposed in this paper, including support vector machine (SVM), picking up Mel frequency cepstrum coefficient (MFCC) characteristics of isolated words to distinguish the speaker's age. The voicebox of this paper includes 4507 isolated word speech. Experimental results show that the recognition rate based on isolated word speech can reach 72.93%. Through the experiment towards SVM classifier, we could find that the performance is improved without normalization for MFCC.

Keywords: Isolated words; Mel frequency cepstrum coefficients; Age recognition; Support vector machine

1 Introduction

The human voice not only provides the semantics of spoken words, it also contains speaker dependent characteristics. Speaker age recognition is an essential technique in automation speech recognition based on the speech waveform parameters in speaker's voice, and it can improve the latter's performance [1]. With the increase of age, not only face but also the voice will change [2], therefore the human voice can be used as an important characteristic to represent the person's age [3]. Speaker age recognition is according to speaker's voice in speech waveform parameters. The voice recognition technology does not need too much contact and interaction patterns between the entities of certification, therefore, it is undoubtedly much more convenient than other identification technologies. During the configuration process, the electronic equipment, on which sound card and microphone have been installed, can save a lot of cost. In addition, as for the use of cases, it does not need high requirements as monitoring equipment, so the usage of environment will be more arbitrary and private.

Most current speaker age recognition systems based on face recognition [4] and voice recognition. Lots of machine learning methods can be applied in the speaker's age estimation, such as Decision Tree Model (DTM), Artificial neural network (ANN), *k*-Nearest Neighbour (*k*-NN), Naive Bayesian Classifier (NBC), Support Vector Machine (SVM), etc. Minematsu *et al*. [5] put forward an automatic age estimation based on phonetic features of the estimator at first, which could judge whether speaker is old. Muller *et al*. [6] proposed an age algorithm using ANN to divide speakers into 4 groups and the overall accuracy was 64.5%. The accuracy of categorizing the speaker for the elderly and not the elderly based on ANN could be as high as 96.57%. Shafran *et al*. [7] showed that standard speech/ nonspeech Hidden Markov Model (HMM) [8], conditioned on speaker traits and evaluated on cesptral and pitch features, which worked well above chance for gender, age, dialect and emotion traits. When using cepstrum feature to estimate the speaker's age, the accuracy was 68.4%. While using cepstrum and fundamental frequency feature for age estimation, the accuracy was increased to 70.2%. Schotz *et al*. [9] classified speaker's age by classification and regression trees, where using 50 acoustic characteristics the best performance was 72.14% and the average error was ± 14.45 years. Experiments showed that when using 78 acoustic features estimation correct rate increased slightly and the average error was about ± 14.07 years.

Nowadays, the speak age estimation methods are mainly based on the classification method [10] such as DTM, ANN, *k*-NN, Naive Bayesian Model and SVM [11]. These algorithms can effectively classify speaker's age when the request is not very rigorous. In spite of this, the study about speaker age identification is limited. It is usually used in the classification of the large span, such as classify speaker into old/non-old people, etc.

This paper adopts extracting Mel frequency cepstrum coefficient (MFCC) features in isolated words and identifying the speaker's age via SVM classifier. We first extract speech feature parameters as characteristic vector of different ages and add class labels to them. Then we train the feature vectors of different ages of normalized so as to get the SVM sets. What is noteworthy is that when making the SVM training we non-normalization on the extracted features. Figure 1 shows the recognition system framework. Experiments on Beijing Speechocean database [12] succeed in dividing the speaker's age into the youth, adults and the elderly so that has validated the superiority of our method.

978-1-4799-4719-5 /14/$31.00 ©2014 IEEE

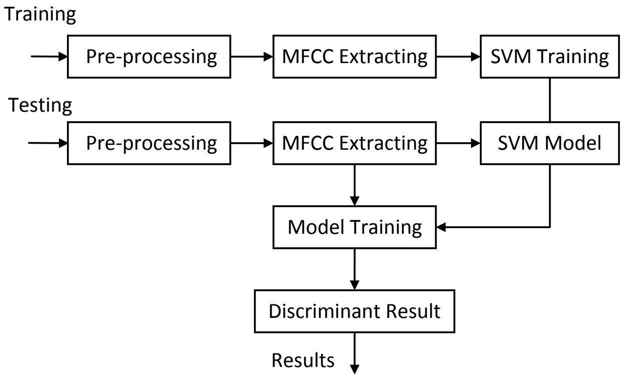

Figure 1 Speaker age recognition framework

The rest of this paper is organized as follows: Section 2 shows the details of our method. Experiments and discussions are shown in Section 3. At last, some conclusions are drawn in Section 4.

2 Pre-processing

The pretreatment of speech signal is also known as the front end processing, which processes the original speech before extracting features in order to make the processed signal better meet the actual needs. It is significant to improve the processing accuracy.

Whether we can take advantage of the speaker's voice to make age discrimination significantly depends on the pretreatment of original audio signals. Pretreating the input audio files can improve the accuracy of feature extraction and recognition algorithm, so preprocessing module is the foundation of the whole system, which mainly includes: endpoint detection, speech enhancement, pre-emphasis, enframing, windowing, etc.

2.1 Endpoint detection

Speech endpoint detection occupies an important position in speech signal proessing. Effective endpoint detection of speech signal is able to analyze and train better, in this way the speech recognition could reach a better performance.

This paper uses two levels of classification endpoint detection method based on the energy and zero crossing rate. Before endpoint detection, the threshold of short-time energy and zero crossing rate [13] should be set. When energy and zero crossing rate beyond the low threshold, the starting point is added into the transition period. During the transition period, since parameter values are relatively small that it is not sure whether it is true speech or not, so as long as the two parameter values are falling below the low threshold, it returns the current mute state. However, if there are two parameters anywhere in the transition section higher than the high threshold, it is sure to enter the speech segment, and mark the starting point. If the current state is in speech, while lower values of the two parameters falls below threshold, and the duration is greater than the set maximum time threshold, it is regarded as voice over, then returns to the parameter value where it is below the lower threshold, and mark the end point.

2.2 Speech enhancement

Nowadays, the recognition rate of speaker age recognition based on speech in the laboratory environment is not as high as we expectations. The denoising methods including spectral subtraction, wiener filtering, etc. In this paper, we adopts the spectral subtraction processing the speech. The principle of spectral subtraction speech enhancement technology is shown in figure 2.

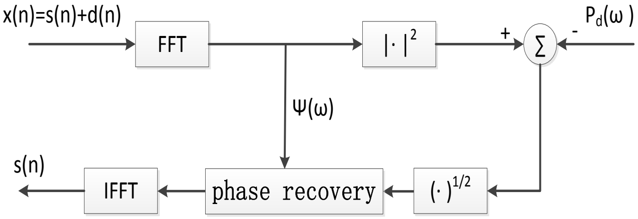

Figure 2 Principle of spectrum subtraction speech enhancement technology

2.3 Pre-emphasis

The voice signal in high frequency part suffers from attenuation. In spectrum analysis, the higher the frequency, the smaller component of that. For speech pre-emphasis it can promote voice high frequency part and make it flat spectrum in order to analysis and process the spectrum much better.

Pre-emphasis generally achieved with 6 *db*/octave digital filter frequency characteristics. When the speech signal is affected by the glottis excitation and nose or mouth radiation the high frequency components will have larger decline therefore it is absolutely essential to through a high pass filter before pre-emphasis. The filter can be expressed as

$$H(Z)=1\text{-}\mu z^{-1} \tag{5}$$

where μ is pre-emphasis coefficient and set to 0.9375.

2.4 Enframing and windowing

Because of the quasi stationary characteristic of speech signal the properties and characterization of the speech signal parameters were changed over time. Only in the short period the speech signal can be considered as a stationary process. In order to decrease Gibbs effect and reduce the slope at the ends of the frame, a hamming window is to prevent the ends of speech frame change sharply to zero. The hamming window expression is shown as follows:

$$W(n)=\begin{cases}0.54-0.46cos\left[\dfrac{2\pi n}{N-1}\right], & 0\le n\le(N-1)\\ 0 & ,\ n=others\end{cases} \tag{6}$$

The waveform multiplied by the hamming window is compressing the close function waveform at the ends of the parts. So even in periodic spectrum analysis of the obvious dullness, multiplied by a suitable window function can also restrain pitch cycle analysis interval change of phase relationship which can obtain stable spectrum.

2.5 MFCC extraction

Mel-Frequency Cepstrum Coefficients are utilized by imitating human auditory mechanisms, because the sound perception level of human ear versus frequency is not a linear proportional relationship. Mel-frequency generally corresponds to the actual frequency of logarithmic distribution relationship. Mel-frequency relationship with actual frequency *f* can be approximated as

$$Mel(f) = 2595 \cdot \lg\left(1 + \frac{f}{700}\right) \tag{7}$$

where the actual frequency *f* is measured in Hz.

Since the critical frequency bandwidth are consistent with the growth of Mel-frequency, we can according to the size of the critical bandwidth from low frequency to high frequency to put a band-pass filter bank, generally choose the triangular filter sequence. Filtering the input signal, then output each filter signal energy as the basic characteristics, through further processing can be seen as the input characteristics of the speech signal. In addition, this feature does not depend on the nature of the signal. There is no limit or hypothesis to the input signal. Therefore, the parameters of the auditory closer to the properties has better recognition performance. The MFCC extraction steps are shown as follows:

Firstly, conducting discrete Fourier transformation (DFT) on the preprocessed sound signal $x(n)$ to get frequency spectrum $X(k)$:

$$X(k) = \sum\nolimits_{n=0}^{N-1} x(n) e^{-j2\pi/N}, 0 \le k \le N \tag{8}$$

where $x(n)$ denotes the input sound signal, N is the length of DFT.

Secondly, we calculate the square of $X(k)$ to get the energy spectrum, i.e.

$$P(k) = X^2(k) \tag{9}$$

Thirdly, filtering the energy spectrum based on Mel triangle filter group and the transformation function of filter is formulated as:

$$H_m(k) = \begin{cases} 0 & , k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) < k < f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) < k < f(m+1) \\ 0 & , k > f(m+1) \end{cases} \tag{10}$$

where $0<m \le M$, M is the number of filters, $f(m)$ is the number of filters, $f(m)$ is the central frequency of each triangular filter. Therefore, the output $S(m)$ of filter group can be described as:

$$S(m) = \ln\left(\sum\nolimits_{k=0}^{N-1} \left|C_\alpha(k)\right|^2 H_m(k)\right), 0 \le m < M \tag{11}$$

Finally, the MFCC can be obtained by discrete cosine transformation (DCT) regarding to $S(m)$ as:

$$C(n) = \sum\nolimits_{m=0}^{N-1} S(m) \cos\left[\frac{\pi n(m-0.5)}{M}\right], 0 \le n < M \tag{12}$$

where $C(n)$ is desirable MFCC, M is the number of filters. In the subsequent experiments, M is set to 24.

2.6 Support vector machine

Support vector machine is a new promising classification proposed by AT&TBell research labs led by Vapnik in 1963. The rapid development and improvement of SVM has many advantages in solving small sample, nonlinear and high dimensional pattern recognition problems. It can be applied to other machine learning problems such as function fitting and rapid development, now in many fields svm has achieved successful applications. In the case of linear separable, SVM constructs a hyperplane *H*

$$\omega \cdot x + b = 0 \tag{13}$$

where ω is weight vector, x is eigenvector and b is parameters.

By solving the quadratic optimization problem can get optimal hyperplane.

$$\begin{gathered} \min \Phi(\omega) = \frac{1}{2}\|\omega\|^2 \\ s.t.\ y_i(\omega \cdot x_i + b) \ge 1, i = 1,2,\dots,n \end{gathered} \tag{14}$$

where x_i represent the i training sample feature vector, y is the sample classification mark.

When the situation is extended to nonlinear add penalty parameters to the wrong data. Get a new function:

$$\begin{gathered} \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{R} \varepsilon_i \\ s.t.\ y_i(\omega^T x_i + b) > 1 - \varepsilon_i, \varepsilon_i \ge 0 \end{gathered} \tag{15}$$

Selecting the proper kernel function and then mapping the feature vector to high-dimensional space to achieve an accurate classification.

3 Experiment and analysis

The experiments in this paper are based on a corpus of speech data collected from Beijing Speechocean. The sampling frequency is 22.05*kHz* (16bit). It was recorded in the quiet (office) environment serving as single-channel recordings that contain 4,507 periods of speech. Additionally, the speakers involved in the recordings have not added any emotional tone. Pronunciation candidates speak mandarin are from 22 provinces, municipalities or autonomous regions in the relative standards. The content in the corpus includes names, addresses, countries of isolated words such as CaoCao, Beijing, Italy, etc.

3.1 Age division

The person's age is divided into the following several intervals by the United World Health Organization, such as 0~44 years old for young people, 45~59 years old for middle-aged, 60~74 years old for old people, 75~89 for the elderly, the age of 90~90+ for longevity. Because of the larger span of young and middle-aged people do not meet the requirement for speaker age recognition. Therefore based on person's physical, mental and going through developmental changes and combined with the traditional Chinese method of age, the age is divided into the following three stages. The three stages are the youth: 12~29 years old, adults: 30~44years old, the elderly: 45~65years old.

3.2 Normalization

Normalization is to limit data which need to be processed in a certain range. Usually involved in the process of experiment data such as speech signal or image signal. In order to analyze the rule of the next step, we will process normalization of data. But based on the experiment we could find out that not all cases are suitable for the normalized processing. As shown in Table I.

Table Ⅰ The normalization results

Normalization or not	Accuracy rate
Yes	68.75%
No	72.93%

The experiment shows that when using the SVM training classification, non-normalized data can further improve the recognition effect by 4.18%. Experimentally, when using this method normalizing the original data, it means has admitted a hypothesis that the maximum of testing data sets is less than the maximum of the training data set. This assumption is obviously too far-fetched, because actual situations are not always necessarily like this.

3.3 Kernel function selection

Kernel function is the core content of the SVM, whether to choose a suitable kernel function has great effect on the results. In general, as for the linear problem it is better to use linear kernel function, while as to nonlinear problems Radial Basis Function (RBF) kernel function may be better. The experimental results of different kernel functions for SVM classifier are shown in Table II.

The experiment shows that when using RBF kernel function, the final classification accuracy rate is best as reaching 72.93%. What's more, the linear kernel function is applied to linear separable case, the experiments belong to the linear inseparable. The RBF kernel function can take a sample mapped to a higher dimensional space, and it needs little parameters than polynomial kernel function and it will reduce numerical computational complexity. As a result, we use RBF kernel function in the subsequent experiments.

Table II The experimental results of different kernel functions

Kernel function	Accuracy rate	Svmtrain parameter
Linear	69.23%	'c 2 g 1 t 0'
Polynomial	67.94%	'c 2 g 1 t 1'
RBF	72.93%	'c 2 g 1 t 2'
Sigomid	62.08%	'c 2 g 1 t 3'

3.4 Further discussions

Here we still carry out some experiments towards speaker age recognition based on isolated words with non-normalization for MFCC and RBF kernel function for SVM classifier.

Figure 3 shows the average age recognition accuracy of male and female, we can see that the recognition accuracy of male is higher than that of female by 2.3% approximately, because the age ambiguity of female is more common, which makes the age recognition more difficult.

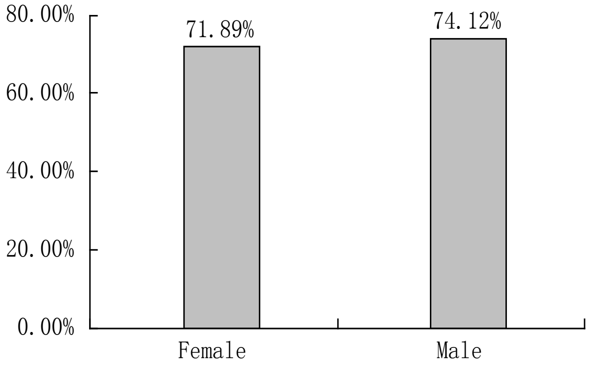

Figure 3 Average recognition accuracy of male and female

Table Ⅲ illustrates the average recognition accuracy of different age periods, which is obtained by averaging the accuracy of both male and female in certain age period. It generally can be concluded that the accurate rate of the age between 45 and 65 is best. This is mainly because in this age period, most people's voice features have been stereotyped and the MFCCs are more stable for this case, which leads to discriminate age easily.

Table Ⅲ Average recognition accuracy of different age periods (Each section contains both male and female)

Age	Accuracy rate
12~29	73.06%
30~44	72.59%
45~65	73.45%

3.5 SVM vs HMM

In this part, we provide two different methods based on SVM and HMM. The results are represented in figure 4.

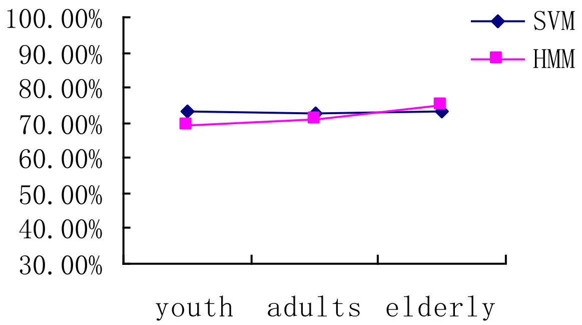

Figure 4 Recognition accuracy of SVM and HMM

Through the experimental analysis we know that HMM is suitable for continuous signal and the results reflect the similarity of similar samples. Nevertheless, the output of the SVM reflects the differences between the heterogeneous samples. Therefore, SVM is more suitable for the classification problems we discussed in this paper.

4 Conclusions

We searched the changes of the respiratory system along with the age for the suitable voice signal characteristics, and then choose an effective classification model, establishes the overall framework and process of the experiment. This paper altogether conducted four experiments. First about the normalization. For this we found as for the SVM classification training, experimental accuracy without normalization is 4.18% higher than normalization. The second is Kernel function selection. There are four kernels which are Linear kernel functions, Polynomial kernel function, Sigomid kernel function and RBF kernel function. Through the contrast, it can be seen that the final classification accuracy rate was the highest when using RBF kernel function. Third, using cross-validation to estimate the ages of each speaker to get the average recognition accuracy. Exclude the speaker's data from all other than training as the training set, then we get recognition accuracy. From the experiment, we obtained the average accuracy is 72.93%, the highest recognition rate is 87.93%, and the lowest recognition rate is 66.72%. The recognition result reached the expected aim. This experiment consider the influence of the noise is less, but due to the real environment may be more complex, therefore for the practical application of this method it is necessary to further explore and research. Our method has a much better effect than the traditional methods which recognition whether the speaker is the old. It obtains a good result and reduces the computational significantly. The result shows that we propose a simple, efficient method to speaker age recognition. Last but not the least, by comparing the experimental test of SVM and HMM we got a result that based on SVM is more appropriate for the article

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046, 60875050, 60675025), National High Technology Researchand Development Program of China (863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730, JCYJ20130331144716089).

References

[1] T. Bocklet, A. Maier, J.G. Bauer and F. Burkhardt, Age and generd recognition for telephone applications based on gmm supervectors and support vector machines, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), vol.4, pp.1605-1608, April 2008.

[2] S. Schotz, Acoustic Analysis of Adult Speaker Age, Speaker Classification IBerl, 2007.

[3] S. A. Xue and D. Deliyski, Effects of aging on selected acoustic voice parameters: Preliminary normative data and educationalimplications, Educational Gerontology, vol.27, no.2, pp.159-168, November 2010.

[4] Y. Fu, G. D. Guo and T. S. Huang, Age synthesis and estimation via faces: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 32, pp.1955-1976, November 2010.

[5] N. Minematsu, M. Sekiguchi and K. Hirose, Performance improvement in estimating subjective agedness with prosodic features, in Proceedings of Speech Prosody, vol 11, pp.507-510, April 2002.

[6] C. Muller, F. Wittig and J. Baus, Exploiting speech for recognizing elderly users to respond to their special needs, in Proceedings of INTERSPEECH 2003 8th European Conference on Speech Communication and Technology, vol 1, pp.1305−1308, September 2003.

[7] I. Shafran, M. Riley and M. Mohri, Voice signatures, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03), vol 6, pp.31-36, November 2003.

[8] L. Ming, K. Han and S. Narayanan, Automatic speaker age and gender recognition using acoustic and prosodic level information fustion, Computer Speech & Language, vol 27, pp.151-167, January 2013.

[9] S. Schotz, Analysis and synthesis of speaker age, Travaux de l'Institut de linguistique de Lund, pp.58-63,2006.

[10] V. Heerden, C. E. Barnard, M. Davel, C. Walt, et al, Combining regression and classification methods for improving automatic speaker age recognition, in Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'10), vol.4, pp.5174-5177, March 2010.

[11] C. S. Ram and R. Ponnusamy, An effective automatic speech emotion recognition for Tamil language using Support Vector Machine, International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), vol.7, pp.19-23, February 2014.

[12] http://www.speechocean.com/cn-ASR-Corpora/Index.html.

[13] R.G. Bachu, S. Kopparthi, B. Adapa and B.D. Barkana Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy, Advanced Techniques in Computing Sciences and Software Engineering, [M], pp.279-282, 2010.