Television station identification system based on speech forensics

Wang wenxiu1
Nanjing institute of technology
Nanjing, China
413018821@qq.com

Zhou yunxia2 Zhou jingkai2 Cao dongjian2
Nanjing institute of technology
Nanjing, China
1205891498@qq.com

Abstract : This paper introduces the principle of speech signal processing, the method and the research status. According to the domestic and foreign relevant speech signal processing technology research [1], at the beginning of this research, we prepare enough files. Presenting a BP neural network based on MFCC recording device identification method, we began to record the audio files to be read by the first using MATLAB, and then for each file for MFCC to extract characteristic parameters [2], and then training representative voice files, establish the BP neural network recording equipment model library, then select the taped speech compared with BP neural network model libraries do, determine testing voice from which Television station.

Key word: Speech processing [3],Television station identification,MFCC,BP,MATLAB

I. INTRODUCTION

With the rapid development of information and technology, the security of information requirements have become more urgent. Here for television signals to the background sound of the security of the information provided certain conditions. As the voice has more advantages than other forms of interaction side, so use this technique as a background sound for the information security also provides a stable protection. With the intention of television being identified, we need identify a specific voice and pre-speech feature extraction television programs or distinguish TV channels. Speech signal not only contains semantic content of information, but also includes information about television. Each television channel of its unique characteristics and frequency characteristics of their information has been reinvented its radio feature, which is the fundamental basis for station identification. Because the radio signal is adjusted to a higher level. Wanting to embed the human voice, which is difficult to be heard, the eavesdropping and other forms of interference greater access to information, the confidentiality of information is better, therefore, TV voice signal analysis will be great meaningful. In previous, no such data for station information and identify patterns, through which radio identification will also have a very bright application prospects.

II THE BASIC PRINCIPLES IN RECOGNITION

MFCC feature extraction, including Pre-emphasis, Sub-frame, Window, DFT / FFT, MEL frequency filter, Log logarithmic energy and DCT Spectrum. Quantitative and after sampling, and then through training and the establishment of BP neural network model library voice[4], voice and last comparison test model library (training) speech feature parameters in order to make a final decision. Mainly uses MATLAB simulation to determine the correctness. In the course of the various parameters of the speech extraction method to do a comparison, the comparison also studied voice at BP neural network model and other methods [5].

II. IMPLEMENTATION PROCESS

A. Introduction to Sampling and Quantization

Signal from the analog-to-digital conversion is divided into two processes sampling and quantization. Sampling is continuous time discretization. Quantization of the sampled signal is converted into a continuous amplitude of a finite set. The sampled signal can be expressed as:

$$S(n) = sa(nT), -\infty < n < +\infty \quad (1)$$

sa: analog signal, n: integer, T: sampling interval[6]

Sampling theorem states that: If the analog signal sa (t) has a Fourier transform limited bandwidth, that is, when the frequency $\omega \geq 2\pi W$ when there sa ($j\omega$) = 0,

then T ≤ 1/2W when the sampling interval, the analog signal fully restored by the sampled signal reconstruction, when the ω is the Nyquist sampling frequency. In the sub-sampling the adjacent high and low frequency spectrum overlap, causing distortion. In order to avoid distortion, should be limited to the bandwidth of the analog input signal, or increasing the sampling frequency, to ensure compliance with the Nyquist sampling theorem. The quantification can be divided into uniform quantization, non-uniform quantization of quantitative, differential quantization and so on.

B. 2.2 Feature Extraction

MFCC parameters are calculated by frame, its extraction process can be Figure 1 shows, because everyone's talking voice lines and channels are different, so we usually use this technique to increase the speech signal analysis, to function as the speech signal after sampling , inserting a plurality of high-pass filter to enhance the signal characteristics we need, then the increase in the signal analysis and processing; Mel filtering effect is a special triangular filter the speech signal amplitude squared values are smoothed; the number of operation has the following features, the first is the dynamic range compression speech spectrum, the second frequency component is not converted to the desired component while filtering unnecessary ingredients; discrete cosine transform is to make up each separate voice component .

The parameters' calculation process:

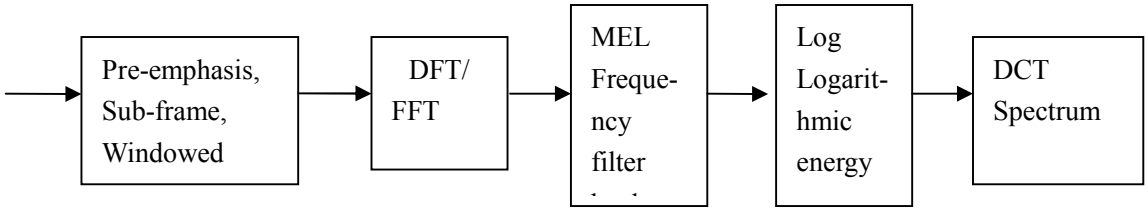

Figure 1 MFCC parameter extraction process

Pre-emphasis: a digital voice signal after sampling y (n) through a high-pass filter:

$$H(z)= 1-a*z^{-1}, 0.9 \leqq a \leqq 1.0 \qquad (2)$$

After Pre-emphasis signal Y (n) is:

$$Y(n)= y(n)-a*y(n-1) \qquad (3)$$

Framing: Sub-frame when the Window borders in order to avoid missing the signal, generally taken as a 10-20ms. When the frame to make the offset, between the frame and the frame part to overlap[9]. To avoid the frame and the large variation in characteristics between the frames, generally the time shifted by one half frame before removing one, which is half the frame length as the frame shift.

Calculate short-term energy: short-term energy represents the size of the sound amplitude, at the volume level, according to the value of this energy to make some minor speech signal noise can be filtered out. We will generally be limited to the energy value below our threshold of a frame as a mute segments.

Windowed: Speech at long range do not deal with, because the voice of the long range is constantly changing, there is no fixed characteristics. Therefore, each frame needs to be substituted into the value of the Window of the Window function is set to 0, the purpose is to prevent the respective ends of the frame signal may cause discontinuities. Commonly used Window function with rectangular Window, Hamming Window and Hanning Window and so on. Window function according to the frequency domain characteristics, we generally use the Hamming Window.

Formula is within the scope of the Windowing

$$\omega(n)=0.54-0.46*\cos(2*pi*n/(n-1)) \qquad (4)$$

Fast Fourier transformation: the time domain, the voice signal changes rapidly and unstable, and the frequency domain, the spectrum of the speech signal

may be slow changes with time, it is generally observed in the frequency domain speech signal up. The frame after the Windowed FFT, the spectrum can be calculated parameters for each frame.

Triangular bandpass filter: The above-obtained spectral parameters of each frame through a set of N bandpass filters triangle (20 ≦ N ≦ 30) which consists of Mel scale filter, the output of each band logarithmic , calculated on the number of each output energy, k = 1,2, ... N. This last parameter cosine transform of N obtained L-order Mel-scale parameters[7].

Calculate the output of each filter bank log energy:

$$S(m) = \ln(\sum_{k=1}^{N-1} | Xa(k) |^2 \ Hm(k)), 0 \le m \le M-1$$

(5)

Where Hm (k) is a special frequency response of the filter triangles.

DCT (discrete cosine transform through): get the MFCC coefficients.

$$C(n) = \sum_{m=0}^{M=1} S(m) \cos(\pi n(m - 0.5/m)), 0 \le n \le N-1$$

(6)

MFCC coefficient number is usually taken 20-30, it represents the spectrum of energy, so in the general recognition system, we call it the energy factor, does not directly say that they are cepstral design selected is 24 order spectrum.

C. 2.3 BP Neural Network

BP (Back Propagation) neural network, that error back propagation error back propagation algorithm learning process forward by the information dissemination and error back propagation of two processes. Artificial neural network must first learn to certain guidelines to learn, and then to work. Network standards for learning should be: If the network make a wrong decision, then through a network of learning, should be making the network to reduce the possibility of making the same mistake next time. First, the network connection weights assigned to each (0,1) interval random value, "A" corresponding to the image mode input to the network[9], the weighted sum of the input mode, compared with the threshold, then non-linear operation, to obtain the output. In this case, the network output is "1" and "0" with a probability of 50%, that is to say completely random. Then if the output is "1" (the result is correct), then the connection weights increased, so that the network is encountered again "A" mode is entered, it is still able to make the right judgments. When the actual output does not match the expected output into the error back propagation phase. Error through the output layer, the error gradient descent by way of correction weights of each layer to the hidden layer, input layer, layer by layer back propagation. Cycle of information dissemination and forward error back propagation process is continuously adjusted weights of each layer process, but also the process of learning and training neural networks, this process continues until the network output error is reduced to an acceptable level, or pre-set given number of learning so far.

III. MATLAB REALIZATION

MFCC function: MFCC parameter extraction procedure used for framing aggravated when other treatment, to provide frequency DCT transformation parameters[8]. In the design, set to 256 samples per frame. The signal x of frame processing, each frame has 128 samples, between adjacent frames overlap ratio of 50%.

Feature extraction radio program: This program will actually two radio recordings divided into 10 segments 60 recordings as a group of six groups, respectively, for training and testing groups GMM parametric modeling. First, the respective recording device to the program for data transfer, then the maximum value of the parameter for normalization, to facilitate the calculation of the actual parameters and observations, and then set to 24 spectrum dimension 256 samples per frame spectrum transform and collect spectrum parameters are written 6 files, create GMM model libraries.

Using MATLAB functions to read audio file comes, the interface design.

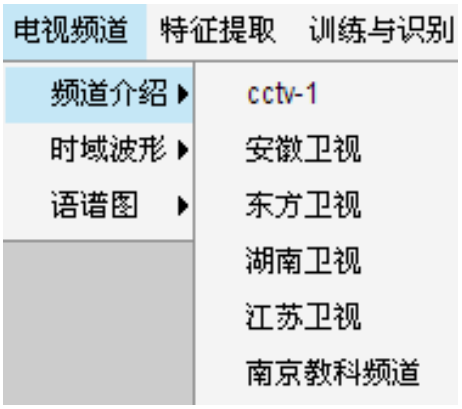

Figure 2 Interface diagram

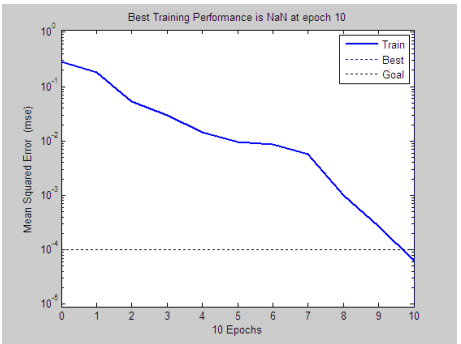

Figure 3 Training convergence

Figure 4 Recognition results

Finally depicts the flowchart of the program design. After changing the parameters of the algorithm and feature comparison, contrast ratio has been greatly improved, can reach 88% recognition rate, remains to be explored further, to find the best parameters of comparison methods to improve the recognition rate.

IV. CONCLUSION

This design is still owning broad prospects in development, such as play the same program for different TV brands, their associated bit rate and sampling frequency and other parameters of different devices can also be identified; same program in different time periods, as well as different recording equipment, sound recording may lead to different parameters affect the device recognizes the success rate, so you can improve as much as possible in the future be used speech database, so you can more effectively carry out the correct identification of stations and their programs; parameter settings requires a lot of experiments trying to analyze the results, in this way we would make the function more accurately and complete functionality.

ACKNOWLEDGMENT

Supported by the National Natural Science Foundation of China (Grant No. 51075068)and supported by Provincial college students training plan(Grant No.201211276113 No.201311276016)

REFERENCES

[1] Phil Rose.Forensic speaker recognition at the beginning of the twenty-first century-an overview and a demonstration[J].Australian Journal of Forensic Sciences,2005.

[2] Bingxi, flexor Dan, PENG Xuan and so on. Practical speech recognition foundation [M]. Beijing: National Defense Industry Press, 2005.1

[3] Chen Liwei. Based on HMM and ANN Chinese Speech Recognition [D]. Harbin: Harbin Engineering University, 2005

[4] Liu Weiguo. MATLAB programming tutorial [M]. Beijing: China Water Power Press, 2005

[5]Ftrm i S Speaker Independent Isolated Word Recognition Using Dynamic Feature of Speech Speetrua1[J].IEEE Trails on AcousOcs,Speech,Sigmd Processing,1986,34(1):52—59

[6]Wayman J L,Reinke R E and Wilson L. High Quality Speech Expansion,Compression,and Noise Filtering Using the SOLA Method of Time Scale Modifieation[A].In 23rd Asilomar Conferenee on signals ,Systems,and ComPuters[3],1989,2:714-717.

[7] A. Paoloni , A.Federico , G.Ibba , " Comparing direct spectral matching techniques whit formant extraction ones for speaker recognition "[J], Proc. of the 12th ICA congress, Toronto , July 1986

[8] G.R. Doddington , "Speaker recognition -Identifying people by their voices "[J] , Proc. of the IEEE , Vol. 73 , No11 , November 1985

[9]D.J. Hejna.Real-Time Time- Seale Modifieation of Speeeh via the synchronized OverlaP-AddAlgorithm[D].Master Thesis,Cambridge University,1990