Dynamic Speech Feature Parameter Extraction Based on Fitting*

Meng Yingjie, Liu Wenjun, Bai Lixin and Chen Wei

School of Information Science & Engineering, Lanzhou University, Lanzhou, China

{mengyj & liuwenjun11}@lzu.edu.cn, bllovy@126.com,cursedvampire@163.com

Abstract - In view of the existing research of the speech feature parameter recognition, the anti noise is poor and storage capacity is larger. So, data fitting has been introduced into speech feature parameter extraction to enhance that. Combine with speech spectrum dynamic changes and the short-time energy smooth stationary of speech signal, this paper puts forward and designs a arithmetic of dynamic speech feature parameter extraction based on fitting, and constructs the feature parameter extraction and personal identification scheme. And also designs critical modules algorithm. The detail process of feature parameter extraction: firstly, it created 2-d coordinate for each frame data. Then, we use 2-d coordinate system to fit for making the fitting function is matched primary data perfectly, and get the best fitting order of each frame. Lastly, it extracts the feature parameter which has been combined with the fitting order in each frame. The arithmetic has been simulated an experiment, in order to confirm the applicability and feasibility. The results illustrates the method has preferable anti-noise performance, especially expression and storage for speech segment feature parameter show more obvious advantages.

Index Terms - speech recognition, feature parameter, extraction method

1. Introduction

Speech signal processing technology and its application have become the indispensable important part of information society; Speech recognition is a very important research side of speech signal processing, and is the important techniques of human-computer interaction. Because of speech signal contains semantic information, personal characteristic information and environmental, etc. Therefore we can analyze and extracting of the speaker personality characteristics, used it for automatic identification of the speaker's identity. Compared with other biometric technologies, the advantage of speech recognition is not lost and forgetten, no need memory, ease to use, etc. There is also a high degree of user acceptance, does not involve privacy users without any mental disorder, etc.

There are many research results about how to accurate separated and extracted the speaker personality characteristics from speech signal. Currently, there are basically the following categories: Linear Prediction Cepstrum Coefficient (LPCC), Mel-Frequency Cepstrum Coefficient (MFCC), Perceptual Linear Predictive (PLP), etc.

About LPCC, in document [1-4], we know it is the repress- entation for linear prediction coefficient in cepstrum domain. The feature is based on the speech signal value of autoregressive signal, using linear prediction analysis to obtained cepstral coefficients. The advantage is the relatively small by calculation, easy to implement, a better description for vowel. The disadvantage is poor description for consonant, anti noise is poor, and no use of human auditory feature.

About MFCC, in document [2,5,6], it is the cepstrum coefficient extract from Mel scale frequency domain. First, spectrum would convert to nonlinear spectral which is based on Mel-frequency standard, then convert into cepstrum domain. This feature parameter is based on hearing mechanism, can better reflect the individual characteristics of the speaker and be able to distinguish between different speakers. MFCC is better reflecting the characteristics of the speech signal than LPCC. But calculation and storage capacity is large, anti noise is also poor and large amount of data can seriously affect the performance of speech recognition systems. For the above advantages and disadvantages of a single parameter, now put forward many improvements, such as MFCC and LPCC[3], the pitch contour characteristics auxiliary MFCC parameter[7], LPCMFCC[8] ,etc. In certain extent, it improved the recognition rate. But the computation and anti noise effect is also poor. For example, document [3] verified the mixing parameters of MFCC and LPCC, improved the stability of the parameters, and the recognition rate is increased than a single parameter, but their robustness and noise immunity performance is still not good, computation and storage capacity is increased.

In summary, because feature extraction of existing research methods are mainly for speech static characteristics, very sensitive to noise, by increasing the feature parameter's order and dimension to improve the recognition rate, only increased the storage and the burden of speech recognition systems, can't really solve the primary problem for feature extraction.

Therefore, based on the above problems, this paper introduces the idea of fitting, making use of speech spectrum dynamics change and speech signal short-time energy smooth stationary, and put forward a based on fitting dynamic speech feature extraction algorithm.

2. Dynamic Speech Feature Parameter Extraction Based on Fitting Model Construction

Speech signal is complex signal and non-smooth, but it's feature is essentially unchanged in a short range and is stationary signals. Therefore we can make use of speech

* This work is partially supported by NSF Grant #2003168 to H. Simpson and CNSF Grant #9972988 to M. King.

signal's properties of short-time stability, combined with the idea of curve fitting to extract best fitting order, then according to speaker's speech feature of dynamic continuous changes, so integrate all the best fitting order together to form feature parameters for identity recognition.

In order to obtain accurate personal speech feature information, the process can be roughly divided into two stages: pretreatment and speech feature extraction. Considering the speech signal's complexity and non-smooth, the main task of pretreatment is reduce speech feature extraction process workload and filter out background noise, etc. It includes sampling and quantization, denoising, preemphasis, framing and adding windows, endpoint detection, etc. Current in this research has been a lot of mature research fruits, for example, in the document [2,11] recount full, so the details are not mentioned here. The work of speech feature parameter extracted is based on the frame and paragraphs data of the frame after pretreatment. The specific steps are as follows: firstly, make the each frame's data points as input data and obtaining 2-d coordinates. Secondly, combine 2-d coordinate of single frame data point with least square method, dynamic selection the order to fitted this frame data point, and get the best fitting order as single frame feature parameters. Then, integrate all the frame's feature parameters together to form the feature parameters of the speech segment signal, and for storage.

From the above, we can get the speech feature extraction based on fitting process model as shown in Fig. 1.

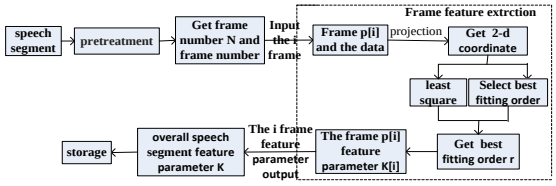

Fig. 1 Speech Feature Extraction Based on Fitting Model

Speech recognition is based on archived speech features database to determine the identity of the person by the detected speech segment. To complete this process, need to extraction feature parameter K' from the to be detected speech segment, its extraction process is similar with the previous feature extraction process. Then calculating the similarity with K' and sample feature set D and for judgment, here the similarity can be calculated by correlation coefficient method. So we can speech recognition model which is based on fitting feature parameters as shown in Fig. 2.

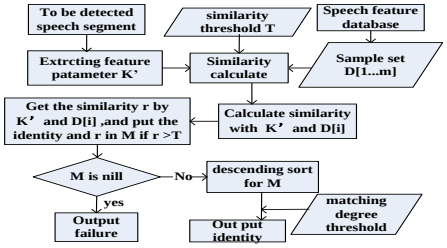

Fig. 2 Speech Recognition Based on Fitting Feature Parameters Model

3. Feature Parameter Extraction and Recognition Algorithms Design

A. *Feature Parameter Extraction Algorithm*

Feature parameter extraction is very important link in personal identification which is based on speech. Based on Figure 1, we know dynamic speech feature parameter extraction based on fitting algorithm mainly composing of single frame's coordinate generation, best fitting order extraction, etc

1) Single Frame's Coordinate Generation Algorithm

Coordinate generation for fitting in single frame's data point, making the fitting curve is best approximation the primary data. So according to single frame's data, we can get coordinate set pw. And get the best fitting order using pw by fitting. Hypothesis Q[1..N, 1..q] represent speech's N frame data set, and N is total frame number and q is the length of frame(q is 256 in this paper); pw is represent 1-dimension coordinate array which is storage the 2-d coordinate of one frame data. The algorithm of single frame coordinate generation executes as following:

PROC Projection (Q[i,1..q],VAR pw)
INPUT：the data set Q[i,1..q] of frame i
OUTPUT: record array pw
Begin
For j←1 To q Do 【 pw[j].x←j; pw[j].y←Q[i,j];】
End

2) Best Fitting Order Extraction Algorithm

In order to make the fitting function is best match primary data, this paper using least square method dynamic select the best fitting order for each frame.

When fitting the data is n order, we select power function $\{1,x,x^2,\ldots,x^n\}$ as function class. So the fitting function is $\varphi(x) = a_0+a_1x+a_2x^2+\ldots+a_nx^n$(n+1<m, m is the number of data points). The data set (x[1..m], y[1..m]) fitting function's undetermined coefficient is $a=(a_o,a_1,\ldots,a_n)$, and the calculation formula is(1).

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \Omega(x,y,n) = \begin{pmatrix} m & \sum_{k=1}^{m} x_k & \cdots & \sum_{k=1}^{m} x_k^n \\ \sum_{k=1}^{m} x_k & \sum_{k=1}^{m} x_k^2 & \cdots & \sum_{k=1}^{m} x_k^{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{k=1}^{m} x_k^n & \sum_{k=1}^{m} x_k^{n+1} & \cdots & \sum_{k=1}^{m} x_k^{2n} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=1}^{m} y_k \\ \sum_{k=1}^{m} x_k y_k \\ \vdots \\ \sum_{k=1}^{m} x_k^n y_k \end{pmatrix} \quad (1)$$

Hypothesis pw is represent 1-dimension array which is storage the 2-d coordinate of one frame i($1\le i\le N$) data. $\psi(x)$ is fitting expression and *sum* is error-sum.the array A[1..10,1..2] used to storage *sum* and corresponding fitting order. r is best fitting order. The algorithm of best fitting order extraction executes as following:

PROC gls(pw, VAR r)
INPUT: coordinate set pw of frame i data
OUTPUT: best fitting order r of frame i
Begin
A[1..10,1..2]←0; //initial array
For i←1 To m Do 【$a\leftarrow\Omega(x,y,i)$; //got the coefficient a when is i order by the calculation formula (1)

$\psi(x) \leftarrow \sum_{j=0}^{i} a_j * x^j$; //the fitting expression

sum←0; n←length(pw);//got the length of pw

$sum \leftarrow \sum_{t=0}^{n} [\psi(pw[t].x) - pw[t].y]^2$;

A[i,1]←*sum*;A[i,2]←j;】//storage sum and the fitting order

Call sort(A[1..10,1..2],1); //ascending sort basis for first row of A

r←A[1,2]; //got the best fitting order

End

3) Dynamic Speech Feature Parameter Extraction Based on Fitting Algorithm

Based on single frame's coordinate generation and best fitting order extraction algorithm, let the best fitting order of frame i ($1 \le i \le N$) as the frame's feature parameter. Then get the speech segment's feature parameter after combined all the frame's best fitting order. The algorithm of feature parameter extraction executes as following:

ROC CurveFit(Q[1..N,1..q],N,VAR K[1..N])
INPUT：total frame number N, data set Q[1..N,1..q]
OUTPUT: speech feature parameter K[1..N]
Begin
For i←1 To N Do
【 Call Projection(Q[i,1..q],pw);//coordinate generation
Call gls (pw,r);//got frame best fitting order r
K[i] ←r; 】
End

B. Speech Recognition Algorithm Design

Speech recognition is based on archived speech features database to determined the identity of the person belongs by the to be detected speech segment. The Fig. 2 shown the speech recognition is based on similarity computation. And combined similar degree value with threshold of similarity and matching degree to screening and identity recognition.

1) Similarity Computation

Similarity computation is mainly calculate the to be detected speech segment feature and archived one speech segment feature parameter similarity value, and make the value as the standard for similarity degree between speech.

The feature parameters seen as 1-dimensional vector, hypthesis vector $x(x_1, x_2,\ldots,x_n)$ represent one's speech feature parameter in speech characteristics database. Vector $y(y_1,y_2,\ldots,y_n)$ represent to be detected speech feature parameter. The simi is the similarity value of x and y. Using correlation coefficient to calculate the similarity value, and the calculation formula is (2).

$$sim(x,y) = \sum_{i=1}^{n} (x[i]*y[i]) / (\sqrt{\sum_{j=1}^{n} x[i]^2} * \sqrt{\sum_{j=1}^{n} x[i]^2}) \qquad (2)$$

The algorithm of similarity computation as following:

PROC Simil (x ,y ,VAR simi)
INPUT： vector $x(x_1,x_2,\ldots,x_n)$ and $y(y_1,y_2,\ldots,y_n)$
OUTPUT： similarity simi
Begin
n←length(x); sum1, sum2, sum3←0;
For j←1 to n DO 【sum←sum1+x[j]*y[j];//calculate $\sum x[j]*y[j]$
sum2←sum2+x[j]*x[j];//calculate $\sum x[j]^2$
sum3←sum3+ y[j]*y[j];】//calculate $y[j]^2$
simi←sum1/sqart (sum2* sum3);//calculate similarity by formula(2)
End

2) Speech Recognition Based on Fitting-feature-parameter Algorithm

Calculate the similarity value of to be detected speech's feature parameter and sample feature parameter, and select the speech segment and add in M when the value is more than similarity threshold, then combine M and matching degree threshold recognition the identity of the speak.

Hypothesis Q'[1..N,1..q] represent to be detected speech's N frames data set. D[1..m] is sample feature parameter set. T is similarity threshold and ω is matching degree threshold. The algorithm of based on fitting-featuer-parameter speech recognition executes as following:

PROC detect(Q'[1..N,1..q],N,D[1..m],T,ω)
INPUT: to be detected speech total frame number N, data set Q'[1..N,1..q], sample feature parameter set D[1..m], threshold T, ω
OUTPUT： speak's identity
Begin
Call CurveFit(Q'[1..N,1..q],N,K'[1..N]);//parameter extraction
j←1; For i←1 To m Do
【 call Simil (K',D[i],simi); //calculate the similarity
If simi>T Then【 M[j].identity←D[i].identity;
M[j].simi←simi; j←j+1;】】
If j>1 Then
【Call Sort(M); //descending sort basis for similarity by M
i ←1;
While M[i].simi≥ω Do 【 print (M[i].identity); i←i+1;】】
else write('this is failure');
End

4. Simulation Verification and Analysis

The next is to evaluate the algorithm from applicability and effectiveness, anti noise, time and space complexity, etc. And comparative analysis the feature parameter extraction algorithm of MFCC and LPCC, and use dynamic time warping to recognition.

Experimental environment: acquisition twenty people spelling 0 to 9 two-pass in closed environment. One is used to extract feature parameter to building sample speech feature parameter database. Other as to be detected speech. Sample speech feature parameter storage as text file, each row contain the man identity information and the ten speech's feature parameter. In addition, the parameter value in speech recognition: similarity threshold T=0.8 and matching degree threshold ω=0.95.

A. Applicability and Effectiveness

Using speech recognition effect to analysis the applicability and effectiveness of feature parameter, and to contrast MFCC and LPCC in document [2,3].The contrast data on quietness as shown in Fig. 3.

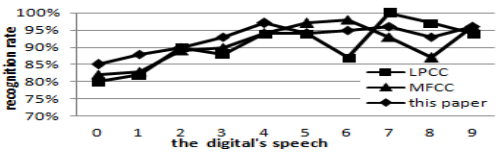

Fig . 3 the recognition effect contrast for three feature parameter

Shown as Figure 3, this paper feature parameter recognition effect is very well, although recognition rate is near with MFCC and LPCC, the recognition rata curve is comparatively stable, reflect the good nature of feature parameter recognition performance

B. Anti Noise

Anti-noise is reflect the insensitivity of speech feature parameter to noise jamming. After plus randomly noise to the to be detected speech, the different feature parameter recognition rata contrast result as shown in Fig. 4.

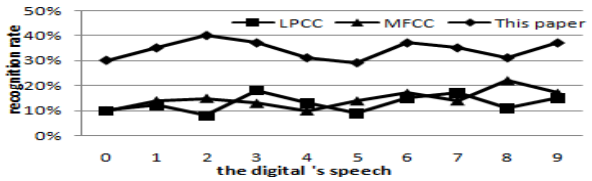

Fig. 4 Plus Noise Recognition Effect contrast for Three Feature Parameter

Although three feature parameter's recognition is very well on quietness. Shown as Figure 4, with the case of noise, MFCC and LPCC recognition rata is very low, it also verified the MFCC and LPCC is sensitivity to noise which is described in introduction. But this feature parameter extraction algorithm is fitting with the data's dynamic trend, a certain degree of flexibility absorb to the data, so the recognition rata is more than MFCC and LPCC in case of noise. Because of plus randomly noise and everyone is different, so affected the results in a certain extent.

C. Parameter's Storage

Sample feature parameter' memory capacitance reflect the size speech feature database's size and impact the recognition system performance. In document [10], we know LPCC and MFCC's feature parameter each component reflect the speech feature's useful information is mainly in before more than a dozen order. And document [2] point out the LPCC's order is 12 good response to speech feature. The document [5] verify the dimension is 24 the MFCC features reflect speech feature. So, LPCC and MFCC parameters for each frame of at least a dozen parameters need to be stored, entire speech segment overall storage capacity more large.

In this paper feature parameter, full use speech continuous dynamic changing, extract one feature parameter to every frame, then the feature parameters of each frame in series, composition can reflect the personality characteristics of speech segments. This paper has a better feature parameters storage benefits.

E. Calculation's Time Complexity

Time complexity reflect the required of computational effort by execution algorithm. According to document [2,5, 10], extracted MFCC and LPCC is need to operating more than ten for each frame. But this algorithm handles up to ten times per frame. This paper feature parameters and MFCC, LPCC's extraction time complexity contrast as shown in Table1.

TABLE 1 Time Complexity

Feature	This Paper	LPCC	MFCC
Time Complexity	O(10n)	>O(10n)	>O(10n)

5. Conclusion

This paper introduces the fitting thoughts, presents a based on speech dynamic changing feature extraction algorithm. This paper algorithm compare with based channel model, auditory model feature extraction algorithm, because of curve fitting is fitting the data dynamic trend of each frame, reflect the data's trend, and increased noise immunity. In the case of adding noise, the recognition rate higher than LPCC and MFCC. Further less feature parameters reduced storage space and improve the efficiency of operation time. Simulation experiment results show that the algorithm has better anti-noise, avai- lability and effectiveness. Because of this paper plus randomly noise and everyone is different, so affected the results in a certain extent, this is the next step for improvement.

References

[1] Wang Biao,Speech recognition system based on LPCC parameter, Electronic Desi gn Engineering, Apr. 2012 vol.20 No.7:18-20.

[2] Zhang Cheng,Researches and Implementation on Speaker Recogn- ition Algorithms and systems, Changsha, National University of Defense Technology, 2005.

[3] Yuan Yujin, Zhao Peihua, Zhou Qun. Research of Speaker Recongnition Based on Combination of LPCC and MFCC. 2010 IEEE International Conference on Intelligent, Computing and Intelligent Systems. xiamen china. IEEE Computer Society :765-767

[4] Xu Fei, Speech Signal Feature Extraction Technology Introduction, the seventh session National Conference of human-computer speech communication collection,2003,45-47

[5] ZHANG Zhen, WANG Huaqing, Improve algorithm of Mel-Frequence Cepstral Coefficients in characteristics extraction based on voice signal, Cmputer Engineering and Applications,2008,vol.44 No.22:54-55

[6] ZHANG Lin, The Resear of Roubust Speech Recognition in Noise Environment Based on MFCC, Hunan University of Science and Technology, 2006.

[7] Shao Yang, Liu Bingzhe, Li Zongge, A Speaker Recognition System Using MFCC Features and Weighted Vector Quantization, Cmputer Engineering and Applications, Dec.2002 Vol.28 No.5:127-128

[8] JIANG Xing-hua, Li Ying, Audio Data Retrieval Method Based on LPC- MFCC, Computer Engineering, June 2009, vol.35 No.11:246-247,253

[9] Zheng Jiming, Wei Guohua, Yang Chunde .Modifiled Local Discriminant Bases and Its Application in Audio Feature Extracion.2009 International Forum on Information Technology and Application. chengdu china. IEEE computer society, 2009: 49-52.

[10] Zhen Bin, Wu Xihong, Liu Zhimin, Chi Huisheng, On the Importance of Components of each Cepstral Components in Speech and Speaker Recognition, Pekinensis Universitatis (Acta Scientiarum Naturalium) , 2001vol.37 No.3:371-378.

[11] Li Peng, Research on Speaker Feature Modeling and Application in Information Security, Xidian University,2008

[12] Ooi Chia Ai, M. HariharanSazali Yaacob, Lim Sin Chee. Classification of speech dysfluencies with MFCC and LPCC features. Expert Systems with Applications. Volume 39, Issue 2, 1 February 2012: 2157–2165