Engineering International Conference 2013 Proceeding
p-ISBN: 97925-2784
November 21st, 2013, Semarang, Indonesia

Cepstrum Parameters for Human Voice Recognition

Susetyo Bagas Bhaskoro
Faculty of Engineering
University of Widyatama
Bandung, Indonesia
susetyo.bagas@widyatama.ac.id

***Abstract*—Several information which often obtained by the human usually through a communication interview. The Human behavior in communicating will produce some ability basic as follow: (1) face to face process, able to analyze something rapidly in remembering type of sound and identity from the speaker; and (2) the process without face to face, able to hear, remember as well as good making decision from identity of the speaker. In this research, make recognition process and identification that previously to be done by the human through quick training process, precise and easy with communicated system. Today the process can be made through computer, started with the process how get characteristic from the sound of speaker and furthermore conclude the result that agree with its training data. The characteristic that found by each the speaker is cepstrum, value of cepstrum that will be stored and made the reference for each speaker on training and examination phases. The examination is made to 12 speakers with the different genders, with each speaker give sound sample as much as 20 times. The examination result of sound recognition process shows the success as big as 70% for Offline examination and 74.167% for real-time examination with total data examination as much as 10 times to each speaker.**

Keywords-sound signal; framing; windowing; fronted detection; cepstrum; extraction fixture; reference fixture

I. INTRODUCTION

The human sound possess different variety different forms but has the same purpose namely to deliver an information such as the translation process result of the message which predicted abstractly in the mind to be delivered to listener with media through the sound. So that the information can be understood with well.

The process for recognition of sound color and frequency from various human sounds that heard will be recorded and categorized through the human brain. Then, the ability of human hearing to recognize from variety sounds very good and not too much the mistake to recognize who has the sound tone.

Through the information advance today, the process to recognize anyone is developed through computation, computer, the human sound is a division from itself (biometric) which can be made as identification. So that the sound previously can be recognized with easy by the human through his hearing, for now the process will be made by computer.

Recognition process of the human sound generally consist of two phases such as: (i) training phase, in this phase can determine some speakers that has the job to give the sound sample, until the system can be stored the reference of sound training data for speaker model, (ii) recognition phase, in this phase must try to agree with examined sound with the model from previously reference storage data and future the decision making through examined sound. If explained in the block like under below.

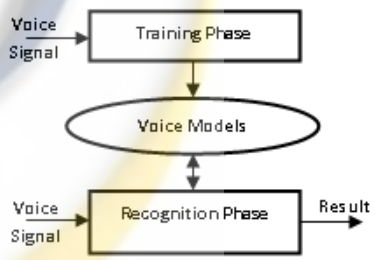


Figure 1. Diagram block of sound recognition

The sound signal is resulted by the human sound and possessing frequency value for each the sound signal. The sound signal which can be made as the sound fixture, and taken the value to be made the basic parameter on codebook.

The sound signal from someone in the training and recognition phases can be the large burden in agreeing it, because occur the big different through the sound when the training and recognition. In fact that the human sound can change as with the time, health condition (fever), noise and used recording media. They can provide the challenge in narrator introduction technology in order to decrease the mistake in making decision through the human sound.

In finishing process of the sound recognition can be explained as follow :

1. Recording of the sound or pronunciation;
2. Preprocessing;

3. Blocking windowing (dividing the signal into frame and decreasing discontinuities);
4. Extracting (each frame, window, frequency);
5. Comparison and Adaption (introduce the pronunciation);
6. Making decision.

II. METHOD

On planning and making, the working follows a diagram block that build and made totally reference material to finish the problem in the final project. The plot of diagram block as follow:

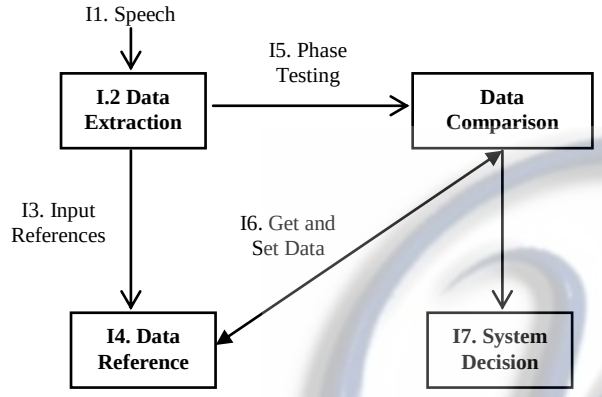


Figure 2. System Design

A. Data Extraction

On the diagram block has function to search fixture values from compared sound for one sound to next sound.

B. Data Reference

On the diagram block has a function to store trained fixtures.

C. Data Comparison

On the diagram block has a function as matching process through the sound fixture that stored with examined sound. System Decision on the diagram block has a function as making decision through suitable sound on previously on the diagram block. So that can identify suitable sound. The diagram block totally from system working as process extending from the diagram block as Fig 3.

III. LITERATURE REVIEW

A. Preprocessing

Before inside to the character taking there is a process namely preprocessing, this process is used to produce outside signal which has the same number of signal. On preprocessing, there are several processes in it, such as: (i) recoding, (ii) sampling, (iii) frontend detection, (iv) normalization. (1) Recording, sound recording can be made on velocity 8000 byte, channel Stereo, time of recording is 1 second and stored with extension, wav; (2) Sampling, follow computation, the resulted

$$Fx \geq f\,max \approx 8000\ Hz \geq (2x3400Hz) \quad (1)$$

Where :

Fs = 8000Hz

Fmax = 3400Hz (2 x 3400Hz = 6800Hz);

(3) Frontend Detection is used to take the sound signal data that consist of pronunciation signal. Then the frontend detection to separate noise signal with other signal like the pronunciation. Usually it is silent noise, on the first and the final of pronunciation, to delete silent (noise), then the frontend detection user is very important; (4) Normalization can be filled with two divisions, such as (i) data long normalization, the normalization has the purpose to add total data so that achieve determined number, (ii) amplitude normalization, it has the purpose to adapt the near distance or range of mouth from microphone on pronunciation time.

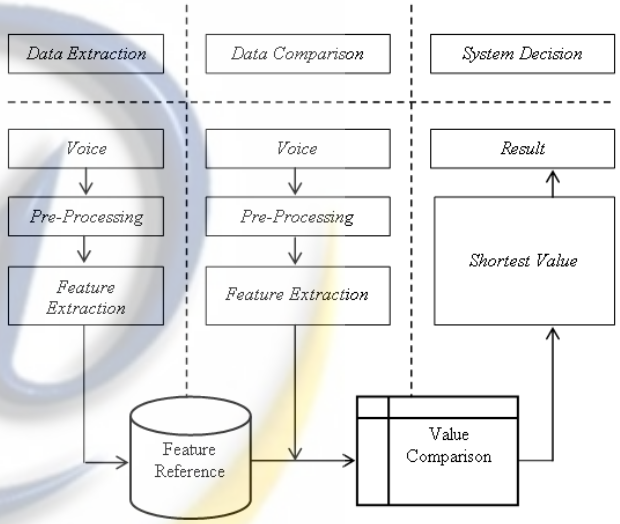


Figure 3. Totally system design.

The characteristic Extraction, in this research, process of signal extraction can be made with using mel-frequency cepstrum coefficient (MFCC) system which produce the characteristic value such as cepstrum, MFCC as used method in managing the sound both introducing and speaker. This method has a main function to follow behavior of human hearing and used to make the characteristic extraction, a conversion process from many parameters[1]. The diagram block from sound extraction uses MFCC method as follow:

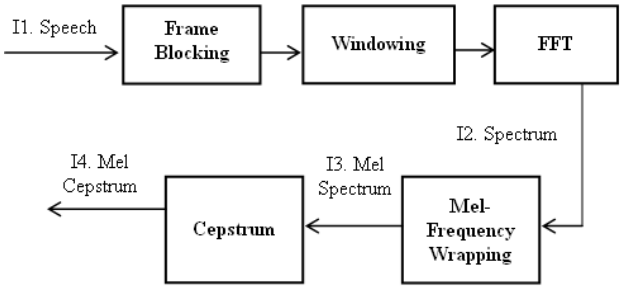


Figure 4. MFCC diagram block

In working process there are explanation from original diagram block, namely the different on diagram block inside, but the result is the same from value characteristic like cepstrum.

According to Alan V. Oppenheim opened Digital Signal Processing come from the paper that publicated by boger at al, which make observation about logarithm on power spectrum of signal. The function from this signal to form the translation from spectrum, because generally, we found the signal that operated to frequency both procedure and conversely. The cepstrum application can be seen as follow:

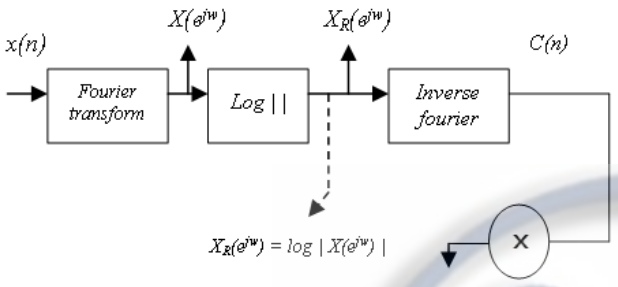


Figure 5. Cepstrum application[2].

Regarding with the different working system from MFCC diagram block, in this case digital signal processing (DSP) can be called homomorphism process that meant the same. In the science and technique it is used to find the difficult signal. The diagram block to make signal extraction can be seen below:

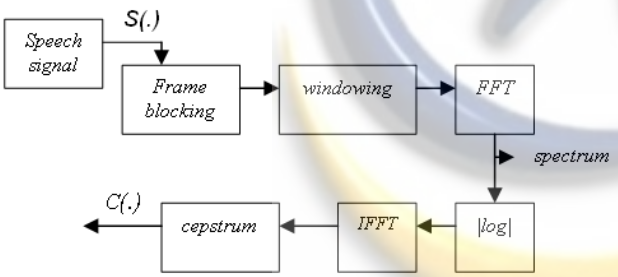


Figure 6. Diagram block of signal extraction.

B. Frame Blocking

Dividing process of sound become several frames that next time can facilitate in computing and sound analysis. Sampling on the final project can be made every time each 20ms, and used sampling frequency as big as 8000Hz, while old recording for 1 second. The number of resulted frame as much as 98 pieces with the number of each frames as much as 160 data.

C. Windowing

Windowing used to delete discontinue, one way to avoid discontinue with window function in order to continue. Windowing function that used in this research is window hamming because hamming function can make the first data and the final frame close 0 value with ell.

D. FFT

Generally a sound signal shown on time domain where we need Fast Fourier Transform (FFT) to change sound signal from tome domain becomes frequency. For sound signal processing, it is very important because frequency domain can be processed with more than time domain. The mathematic equation can be used as follow:

$$\bar{x}(k) = \sum_{n=0}^{n-1} x(n) \,.\, \varepsilon^{-j\frac{2\pi}{n}\underline{kn}}, \quad \varepsilon \approx w_n^{kn} \qquad (2)$$

E. Cepstrum

Cepstral coefficient (ceptrum) is the main result from signal extraction of the sound. Process of FFT that results cepstrum value must pass several diagram block before FFT process. There are 8 data that next time used as the fixture that can present each frame. After data from cepstrum value like numerical is stored in codebook to be made sound long computation.

IV. Our Proposed

Results of a review of the the book which we have done, hence in this study we propose an alternative block diagram using MFCC method. Block diagram of the mel-frequency-wrapping the inside using filter-band as shown in figure 4, we change into several steps as shown in figure 5 and 6.

The goal is to facilitate translation the digital signal processing formulas into a programming language.

V. Recognition Design

A. Vector Quantization

After obtaining sound fixture which pass the training process, then next phase is making the sound comparison that will be examined, furthermore can be concluded that the sound is matching with previously. Vector Quantization is a process to map vector from the space is very wide. Each the region with cluster can be presented by the center that called codeword. The collection from codeword is codebook. Vector Quantization is the training strategy without supervised, in this case is very important to produce the pattern. The speaker with smallest distortion value from codebook can be identified as speaker.

B. Euclidean Distance

For identification process with distortion range from two vector collection that based on the measurement. Euclidean range inter point that measured with a regulation which can be shown by Pythagoreans. Used equation to calculate Euclidean range can be identified with the range among two points $A=(a_1,a_2,a_3,...,a_n)$ and $B=(b_1,b_2,b_3,...,b_n)$.

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2+...+(a_n-b_n)^2} = \sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \qquad (3)$$

VI. APPLICATION TESTING

A. Extraction Fixture

After passing several the diagram blocck phase as mentioned above on sound process, cause some fixturres or called feature vector. The next level is storaging such charactteristic value into database. Used database on this final job is storaging in aplication notepad. Thuss, all vaalues can be stored on appplication.

File result directly stored intto featurre vector for sound type as determined. Used word as much as 1, and pronouncer as much as 12 persons with sepaarating sex become 6 for man and woman. Tottally word that recorded will obtain the same value as the result previously, the number of data as much as 784 data for each model. It will be compared with other data to get a decision that the data not different with storaging media.

For example, examined ssound with some sound refference that owned as follow : tried word model is "bagas" with pronouncer named "bagas", then the next level will be compared with each data. After all feature obtained, next level is compaaring with new sound as examinaation will haave two differents, such as (i) The examination through sound file that recorded prreviously and (ii) Examined soundd file Realtime by pronouncer as registered.

B. Sound Recognition

In the process of examination new sound with comparing stored sound feature previously. It is like with characteristic extraction namely obtain cepstrum value as feature for new sound that not sstored in database. Next the sound feature is compared with each the sound feature that stored.

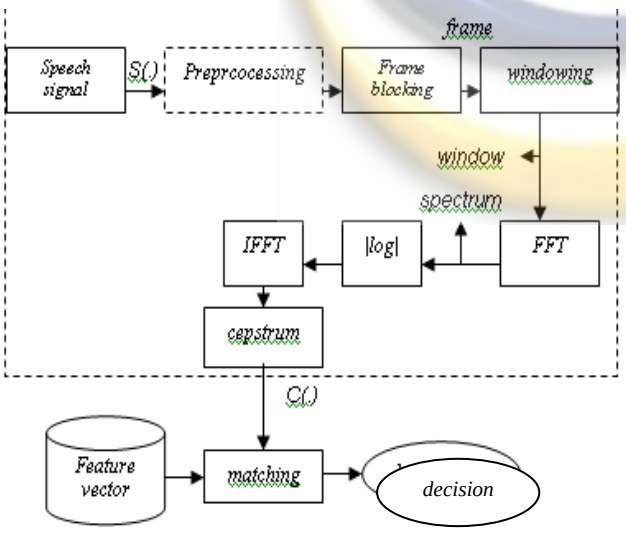
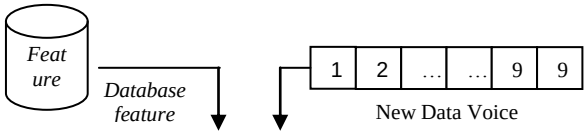


Figure 7. Diagram Block of Sound Characteristic

Procedure from this process as follow the new signal entered into or examined sound will pass the same process after obtained the result of cepstrum value, it must be stored in feature vector but directly compared with stored feature value.



ndex	Number of Data											
	1	2	3	4	5	...	...	94	95	96	97	98
a												
b												
c												
d												
e												
f												
…												

Figure 8. Feature Vector

Procedure on examination coloum, the owned value will be compared one by one into owned sound data. If compared all words, then eeach coloum on sound data will be examined. This process will be used on total examination nameely offline and Realtime, only the xamination Offline previously must be enttered sound file recorded and stored to be ussed on file examination. While Realtime examination, the ssound not owned by file previously, enetered value on the examination file is the value from recorrding in using application.

Stored values in database will be recalleed and entered into charakteristic vector. The number of vector depending on stored registration in database, and process of comparison nameli the new sound. Furthermore the signal fitur enter to application.

VII. RESULT

A. Offline Examination Data

Table below is data from total speaker that examined offline, the result of introduction and failure as well as value range can be known with clearly. The examinaation through sppeaker will be examineed which stored on 10 file. The examined data, the result as follow :

TABLE I. OFFLINE EXAMINATION

Gender	Recognize	Not Recognize	%
Man 1	7	3	70%
Man 2	10	0	100%
Man 3	8	2	80%
Man 4	8	2	80%
Man 5	5	5	50%
Man 6	7	3	70%
Woman 1	6	4	60%
Woman 2	6	4	60%
Woman 3	6	4	60%
Woman 4	5	5	50%
Woman 5	9	1	90%
Woman 6	7	3	70%
Result	**84**	**36**	**70%**

Obtained data totally on table, then result precenttage value from sound examination of offline speeakeer. The number of sound file that examined as much as 120 pieces, from 12 speakers and each they are examineed as much as 10 file. Then, obtained value :

$$\frac{84}{120} \times 100\% = 70\,\%$$

B. Realtime Examination Data

On this realtime examination, the proceess with fulfilling speeaker sound directly into sound application, it is adapted with intonation from each speaker. The result will present identity of speaker that made the examination ddata.

Tabel below is data through totally examined speaker Realtime, the result is :

TABLE II. REAL-TIME EXAMINATION

Gender	Recognize	Not Recognize	%
Man 1	5	5	50%
Man 2	8	2	80%
Man 3	8	2	80%
Man 4	6	4	60%
Man 5	9	1	90%
Man 6	8	2	80%
Woman 1	5	5	50%
Woman 2	10	0	100%
Woman 3	6	4	60%
Woman 4	7	3	70%
Woman 5	9	1	90%
Woman 6	7	3	70%
Result	**89**	**31**	**74.167%**

Data result obtained totally on table, then Stored values in database will be recalleed and entered into charakteristic vector. The number of vector depending on stored registration in database, and process of comparison nameli the new sound. Furthermore the signal fitur enter to application.

$$\frac{89}{120} \times 100\% = 74.167\ \%$$

VIII. CONCLUSION

The working of sound recognition gradually divided into three, such as (i) sound recording, (ii) Extraction of sound, It uses MFCC that results outside value and (iii) sound recognition, after has some fitur data from extraction, the next process will categorize the data become data collection. It is categorized with using algorithm, the final result is shortest from assumed comparison as original file, and then the pronunciation in sound recording with plat tone or neutral, so that it cannot minimize the sound character. The pronunciation with different intonation can decrease the success, although pronounced by the different speaker. Finally, the success on sound recognition offline is 70% for examination as much as 10 times.

REFERENCES

[1] _____,http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition, 2012.

[2] Oppenheim, Alan V. and Ronald W. Schafer, Digital Signal Processing. Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[3] http://www.dspguide.com/. The scientific and engineer's guide to digital signal processing by Steven W. Smith, Ph.D.

[4] _____, http://www.telecom.tuc.gr/~ntsourak/tutorial_sr.htm, 2012.

[5] _____, http://www.willamette.edu/~gorr/classes/competitive.html, 2012.

[6] _____, http://svr-www.eng.cam.ac.uk/comp.speech/, 2012.

[7] _____, http://www.mathworks.com/, 2012.

[8] Proakis, John G. dan Dimitris G. Manolakis, Digital Signal Processing Principles, Algorithms, and Applications. Macmillan, New York, 1992.