2016 Third International Conference on Computing Measurement Control and Sensor Network

Speech Recognition in a Multi-speaker Environment by using Hidden Markov Model and Mel-frequency Approach

Junzo WATADA [1], *IEEE*, Hanayuki [2], ,

[1,2] Graduate School of Information, Production & Systems, Waseda University,
2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135 JAPAN
E-mails: mekymeky@moegi.waseda.jp; junzow@osb.att.ne.jp,

***Abstract*—The sound is a useful and versatile form of communication, where each sound have characteristics and levels of different frequency. Sound serves two basic functions for people around the world: signaling and communication. Several problems are found in sounds identifying, like pitch, velocity, and accuracy of processing voice data. The motivation of this research is to recognize and analyze human voice in a multi-speaker environment from the meeting or indirect conversation. In this research, a Hidden Markov Model approach is proposed as an emotion classifier to carry out testing phases using speech data.**

***Keywords*—Speech Recognition, Multi-speaker Environment, Hidden Markov Model, Mel-frequency**

I. Introduction

A. Research Background

The voice consists of sound made by humans with vocal cords to sing, talk, cry etc. Sound can also be used to make beautiful music through singing, and identify the people we know. Voice allows people to be able to assist in the recognition of emotion, and communicate verbally. There are so many factors that affect a person's voice, such as the length and thickness of the vocal cords. The pitch of someone's voice is one of the factor that human voice different and that can also affected by emotions, inflection and moods. The frightened or excited condition can making the pitch higher, the muscles around the voice box unconsciously contract and putting strain on the vocal cords.

A change in pitch is known as inflection and humans exercise this naturally all the time. To refraining from the screaming, sometimes people also tend to exercise conscious control of the pitch of their voice, because straining and tightening their vocal cords or also changing the pitch of the voice are to mimic someone. Pitch is one of the most obviously biological differences between women and men's speeches. These findings indicate that biological difference in sex can be exaggerated based on social expectations. The social expectations are part of a concept known as gender performativity. Theory about gender is bolstered by continuous performance of individual acts that is a term coined by Judith Butler. A man may perform his gender by wearing tie, or a woman may perform her gender by wearing makeup, skirt or dress, for instance. Some certain aspects change based in culture, pitch of speech also can be symbolic of gender and major factor in linguistic gender identity. But it is not the only one. In the meeting, a group of people present opinions, share information, and make decisions [1]. If difficult to control, the meeting sometimes makes it difficult to achieve the objective. The participants in the meeting may also ill informed, which can lead disagreements and misunderstanding. Consequently, technology that is capable of understanding speech and automatically recognizing used in meetings condition has been attracting and increasing attention as a way to overcome such as problem [2] [3]. In this research, we will explain more features that are different in the human speech.

Corresponding author (Junzo WATADA, Graduate School of Information, Production & Systems, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135 JAPAN, E-mails: junzow@osb.att.ne.jp).

B. Motivation

These days, speech analysis has been widely studied and developed for identifying the voice and emotions behind the speech. Speech recognition has many applications, especially for human machine interaction (HMI) such as voice home caring robots and customer service in call center. Using the voice data we can get many way to develop new technology, and the technologies have been developed for future use and development. So, in this research we want to analyze human voice from the meeting condition or indirect conversation.

This research is to recognize and analyze human voice in a multi-speaker environment from the meeting or indirect conversation. The recorded voice and processed voice can be useful information to recognize gender and language.

C. Research Objective

This research proposes to analyze the pitch and human voice in the quiet place such as meeting place or

978-1-5090-1093-6/16 $31.00 © 2016 IEEE
DOI 10.1109/CMCSN.2016.47

also indirect conversation/call. An Hidden Markov Model (HMM) approach is proposed as an classifier to carry out testing phases using speech data. Left to right HMM will be constructed from signal data Mel Frequency Cepstral Coefficient (MFCC) to recognize the voice. The target of this research is to recognize and analyze the pitch and human voice in a multi speaker environtment such as meeting place or also indirect conversation/call.

D. Outline of the Thesis

This paper is divided into four sections. This Section I briefly introduces the research background, motivation and the research objective of the research. The Section II presents the literature review of research proposed method, review about past studies of voice analysis and about past studies of MFCC and HMM. Section III proposed a experiment for the voice and pitch analysis. Finally, the conclusion and future work of this research is given in Section IV. The direction and mission of future research is proposed based on the brief summary of the current work.

II. Literature Review

A. Overview of Research Proposed Method

The topic of voice analysis or speech recognition is very useful in environment and many applications in our daily life. Speech analysis has been widely studied and developed for many technologies such as text to speech, speech to text, emotion recognition and etc. Several problems are found in speech analysis especially about voice recognition based on gender. This problem usually happened in some conditions, like indirect conversation, call. It is quite difficult to recognize the gender of speaker if the research only focuses in text to speech or emotion recognition.

In this research, we analyze the human voice such as pitch and voice recognition from female and male. The genetic algorithm is used for the training of Hidden Markov Model (HMM) to improve the speech analysis rate in noise environmental conditions. The Mel Frequency Cepstral Coefficients (MFCC) also can be used along with HMM.

Figure 1 below showed the propose method of building a voice analysis system of human pitch based on gender.

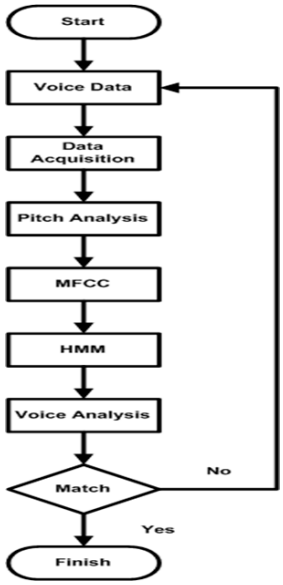

Fig. 1: Figure 2.3 Experiment Flowchart

B. Past studies of Voice Analysis

Voice analysis has received attention recently both in the marketing/advertising literature and among marketing research practitioners. Study about voice analysis, is to provide information about a human speaker. The speech signal is one of the most natural and the fastest method of communication between humans. The speech recognition has wide range in many applications, such as security systmes, telephony military, health-care and also the equipment designed for handicapped [7]. The speech recognition is a promising technology that may radically change the devices that have computational abilities and also change the interface between humans and computers.

The developed applications of speech recognition and voice analysis are very useful such as :

- being used to recognize sentences spoken
- breaking the sentences down into individual words
- Interpreting the meaning of the sentences or words
- Acting on the interpretation in an appropriate manner such as send information back to the user

Some voice analysis software in the recognition part is a tool to judge the emotional state from human voices. It is used in a call center and for a diagnosis of depression mainly. It analyses the emotion of human voice using some parameters which is obtained by data with high speed sampling [5]. Speech to text analysis also one of the research from human voice. Most of speech analytics solutions rely upon phonetic conversion engines or speech to text which are leading to many oversights of meaning, not accurate enough for successful determination of phrases, many false positives and compliance violations. In addition, the adoption of digital channels puts more emphasis on the requirement to text orientations like chat, email, social media interactions, and also to analyze the voice [6].

C. Past Studies on MFCC and HMM

Mel frequency cepstral coefficients (MFCC) are the coefficients that collectively represent the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [6]. One of the example research, studies is about Hijaiyah recognition by using MFCC. By utilizing this system, the expected role of a mentor in introducing and correct

pronunciation of the letters can be replaced so hijaiyah learning to read letters hijaiyah do more independently.

By utilizing this system, the role of a mentor in introducing correct pronunciation of the Hijaiyah letters can be replaced so hijaiyah learning process can be performed more independently. Hijaiyah letter recognition problems can be solved by using the Mel Frequency Cepstral Coefficients (MFCC) for extracting features of voice signal and Hidden Markov Model (HMM) for building voice and performing voice classification [4]. Beside MFCC, DTW (Dynamic Time Warping) is one of feature extraction that can be used for voice analysis. MFCC and DTW are two algorithms for pattern matching respectively and also for future extraction [5].

A HMM is a statistical model for sequences of discrete symbols. This HMM is used and useful for many years in speech recognition and also perfect for the gene finding task. HMM provides a simple and effective framework for modeling time-varying spectral vector sequences. To well known biological problem, HMM also can be applied efficiently. So in bio-informatics HMM gained the popularity, and is used for a variety of biological problems like:

- Protein secondary structure recognition
- Multiple sequence alignment
- Gene finding

One of the problems that sre suitable for the application of HMM is categorizing nucleotides within a genomic sequence, that can be interpreted as a classification problem with a set of ordered observations that possess hidden structure.

III. Experiment and Evaluation

A. Experiment Plan and Objective

The experiment is setted in a silent room or low noise. The voices from 10 female and male participants were recorded. The languages are in English, Japanese and Indonesian. We recorded speech directly from recorder, but we also tried to record the speech through phone line and cellular phone. To avoid experiment affect by noise, the voice recorded in silent room. We collected the data of low noise environment for mixing of these voices through the following software.

Software :

- Sony recorder ICD-SX713
- Microsoft windows 8
- PRAAT
- Matlab R2012b with Signal Processing Toolbox, HMM Toolbox
- Sampling frequency 8000Hz
- 16 bit / sample

B. Data Acquisition

In order to get the good data, the Sony recorder ICD-SX713 was used in this experiment which represented as in Figure 2. The device was a stereo recorder with 2-way adjustable microphones.

The data were processed by using PRAAT software with the 8000 Hz sampling frequency setting.

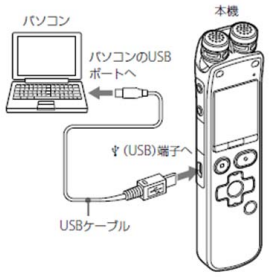

Fig. 2: Diagram of basic recorder tools

C. Experiment Method

Sampling frequency 8000 Hz with the wave format, 16 bit/sample. The acquisition of data samples : $X = Fs.dt$ (second) . (bit/8) . j

where :

X = Signal sampling

Fs = Freq Sampling

dt = Duration audio

bit = bit resolution

$j = 1$ for mono and 2 for stereo

So : 8000 . 2 . (16/8) . 1 = 32.000 byte

To calculate sample rate we used Fs/Ts

Sample rate $= Fs/Ts = 32.000/2 = 16.000$ Hz

The sample point/ frame was gotten with the time sampling data 20 ms

Sample point = Sample rate . Time sampling data

= 16.000 . 0,02

= 320 Sample Point

Praat pitch parameters in this research were set at a minimum 100 Hz and maximum 600 Hz for women's voices, a minimum 50Hz and maximum 300 Hz for men's voices. The results are as shown in the tables below :

In the result obtain by the experiment, we can see that the recorded voice in English and Bahasa Indonesia, the men have lower-pitched voices than women. The female voice in English, the highest is 356 Hz and 140 Hz is the lowest, male voice in English 262Hz is the highest and 77 Hz. The female voice in Bahasa Indonesia, the highest is 277Hz and 152 is the lowest, male voice in Bahasa Indonesia 163Hz is the highest and 96 is the lowest.

D. Experiment Result

The frame blocking in speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector.

$TimeTs = 20ms$, and Sample rate = 16000 Hz,

so the Frame size (N) :

$N = 16000 * 0,02 = 320$ Sample Point

And, $M = 160$

Total Frame : $((I-N)M)+1 = ((16000-320)/160))+1 = 99$

For the Windowing : $w(n) = 0.54 - 0.46cos((2.3,14.0)/(320-1)) = 0.08$ This is formula for converting from frequency to mel scale :

$$\begin{aligned} mel(f) &= (2595 * log10(1+f/700))/(Si/2) \\ &= (2595 * log10(1+0/700))/(0.45/2) \\ &= 4400,37 \end{aligned}$$

So, $Y[i] = 0.4545 * 4400.37 = 1999,96$

We separate all speakers into two subsets for HMM . The first, larger, one is used for compiling and training HMMs, whereas a second subset is used for models testing. We refer those sets as learning and testing. Next waveforms for each recorded word are transformed into characteristic features, which will be referred to as observed sequence, related codebook vector is found.

If the noise contained in the sound is large, the cleaning process of the signal can not be run optimally, because the system can not distinguish between the sound from the environment.

TABLE I: Speech Recognition rate in Bahasa Indonesia

	Jalan	Trunojoyo	Selong	Kebayoran	Baru
FM_1	88.3%	89.9%	89.7%	88.9%	90.3%
FM_2	87.4%	85.6%	85.5%	88.8%	89.3%
FM_3	90.2%	87.5%	87.3%	91.2%	88.9%
M_1	88.6%	86.8%	90.4%	86.9%	88.3%
M_2	89.3%	89.3%	91.1%	86.4%	89.2%
M_3	90.1%	88.7%	89.1%	88.6%	89.5%

TABLE II: Speech Recognition rate in English

	Okay	I	Can	Help	You	With	That
FM_1	93.2%	90.7%	92.4%	89.4%	89.9%	88.3%	90.4%
FM_2	92.1%	91.1%	89.3%	88.3%	89.7%	89.7%	91.3%
FM_3	91.2%	89.7%	88.4%	89.7%	90.1%	90.4%	89.5%
M_1	89.8%	90.4%	90.5%	87.3%	89.3%	89.4%	90.2%
M_2	90.2%	88.9%	89.7%	88.4%	91.5%	88.3%	90.4%
M_3	89.7%	87.3%	88.6%	86.4%	87.7%	89.7%	89.75%

TABLE III: Multi-speaker rate in English

	Recognize	Error
FM_1	73.2	26.8
FM_2	72.6	27.4
FM_3	72.9	27.1
M_1	70.6	29.4
M_2	69.7	30.3
M_3	70.3	29.7

While one speaker detection remains the central problem of speech recognition, there is significant interest in multi-speaker research. The research is clearly essentially harder than one speaker detection.

IV. Conclusions and Future Work

A. Conclusion

Voice recognition process is sensitive to noise because it can affect the voice signal feature extraction process. MFCC method is a good application for feature extraction in speech to catch characteristic which is very important for speech recognition, generating a minimum of data and replicating the human auditory organ in conducting the perception of the sound signal. Based on the introduced method proved a rather reasonable model for the speech recognition, pitch and voice analysis in multi-speaker environment. The human pitch lies in the interval 80-350 Hz, where the pitch for men is normally around 150 Hz, for women around 250 Hz, the pitch is needed to construct this part of the speech signal. In this research we can see that the averages of male voice is lower than female voice. To get the sample point/ frame with the time sampling data 20 ms. By using the method and application in this research we can get the speech recognition accuracy quite high which is more than 85%. The speaker detection remains the central problem of speech recognition, there is significant interest in multi-speaker research. The research are clearly essentially harder than one speaker detection.

B. Future Work

Combination of techniques in parallel processing seems foreseeable with today's fast computers. For the future work we aim the integration of other modalities such as video based or manual interaction will be investigated further. New experiment may be applied for specific analysis in the speech recognition.

References

[1] T.Hori, S. Haraki, T. Nakatani, and A. Nakamura, "Advances in multi-speaker conversational speech recognition and understanding," Feature Articles: Front-line of speech, language, and Hearing research for Heartfelt Communications, pp.1-9.

[2] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," Proc. of the 10th International Conference on Multimodal Interfaces (ICMI 2008), pp. 257–264, Chania, Greece.

[3] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 2, pp. 499–513, 2012.

[4] R. M. Fauzi, Adiwijaya, W. Maharani, "Pengenalan ucapan huruf hijaiyah menggunakan mel frequency cepstral coefficients (MFCC) dan hidden markov madol (HMM) ", Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom. [6] Anjali Bala et al, "Voice command recognition system based on MFCC and DTW," International Journal of Engineering Science and Technology Vol. 2 (12), pp. 7335-7342 , 2010.

[5] B.J. Mohan, and N.R. Babu "Speech recognition using MFCC and DTW", IEEE Advances in Electrical Engineering (ICAEE), International Conference in Vellore, january, 9-11, pp. 1-4, 2014.

[6] Speech and Text analysis retrieved on December 23rd, 2015 from, http://www.genesys.com/platform-services/workforce-optimization/speech-text-analytics.