2011 3rd International Conference on Information and Financial Engineering
IPEDR vol.12 (2011) © (2011) IACSIT Press, Singapore

Probabilistic Speaker Identification to a Non-Continuous Text Independent Verbal Communication

Montenegro , Naparota , Costanilla , Rada
College of Computer Studies, Silliman University, Dumaguete City, Philippines

Abstract. The use of personal features, unique to all human beings to identify or to verify a person's identity is a field being actively researched. Automatic speaker recognition (ASR) is among these. Speaker recognition is the process of automatically recognizing the identity of the speaker on the basis of information obtained from the characteristics of his/her speech.

This project identifies a speaker by method of matching directly the feature patterns of one's voice to the training set. Non-continuous and text-independent recordings of each voice forming the training set are analyzed using Silent Detection and Removal algorithm to separate voiced from unvoiced signals. Linear Predictive Coding in the 10th order is applied to extract the features of the voice and distance is computed using Euclidian Distance Algorithm. The use of Baye's algorithm helps decides the most likely identity of a speaker. The test results of the application yield an average recognition rate of 33.5%.

Keywords: digital signal processing, speech recognition, speaker recognition

1. Introduction

Today's technology provides multimedia devices that not only allow us to communicate but also provide features that inculcate recording of voices. Because of its rampant supply, recording of one's voice became ordinary and in most cases, we take for granted the power of what it contains. Unique to all human being, our voice contains identity features that bind us to who we are. Automatic Speaker Recognition (ASR), a field being actively researched, uses a machine to recognize a person from a spoken phrase. The automatic recognition of a persons' identity is done thru evaluation of the information obtained from ones' speech waves.

There are a number of considerations that should be considered in the evaluation of speech signals to identify speakers. One of those is the method of accepting either a text-dependent or text-independent voice recordings. The former requires the speaker to speak the key words or sentences having the same text, whereas the latter does not rely on a specific text being spoken. Another method is to depict what types of utterances to recognize; non-continuous or continuous. Non-continuous speaker recognition processes voice recordings where separation of each word is emphasized by a short pause or a full stop. On the other hand, continuous speaker recognition allows users to speak almost naturally, while the computer determines the content [8].

The complexity of studying audio signals, understanding its characteristics, is what challenges the group to pursue study on this area. Inspired by the vast components of its application, ranging from voice dialing,

banking, security control to remote access to computers, a small taste of its capability will be studied and implemented by a system that processes non-continuous and text independent speeches [2].

2. Methodology

2.1. System Flow

The user starts the application by inputting voice sample in a wave file format. The voice sample then goes to the pre-emphasis stages that normalize the signal, detects and separates voiced from unvoiced signal. The result then goes to frame blocking and windowing technique wherein feature of the voice is extracted using LPC analysis in the 10^{th} order. A distance computation between the codebook and the test patterns is then computed. Baye's Algorithm (Maximum Likelihood (ML) Formula) is applied to help decide the highest percentage of a match.

2.2. Framing

The input signal is too big for the application to process at once. One way to address this is to divide the signal into frames. The pre-emphasized signal is blocked into frames of N samples, with adjacent frames being separated by M samples. The researchers have chosen 240 samples as the value for N or samples per frame. This is approximately 30 msec. Typically, the value of M is 1/3 of N which is given 82 samples as the value. This will yield 158 samples for the overlapping samples between frames which is approximately 20 msec [5].

2.3. Pre-emphasis and Silence Detection

Not all voices are recorded at exactly the same level, it is then important to normalize the amplitude of each sample in order to ensure that features will be comparable. The procedure is described as:

find the maximum amplitude in the sample, and then scale the sample by dividing each point by this maximum [7].

Since the feature of a voice can only be extracted when there is voice, silence removal technique is applied with the presumption that the silence threshold value is equal to the average value of a normalized room noise recording. The output is a signal of pure word utterances which is normally shorter than the original signal.

2.4. Feature Extraction

One method of feature extraction is Linear Predictive Coding (LPC) analysis. LPC evaluates windowed sections of input speech waveforms and determines a set of coefficients approximating the amplitude vs. frequency function. LPC coefficient is use to separate a speech signals into two parts: the transfer function (which contains the vocal quality) and the excitation (which contains the pitch and the sound). Original speech signal S(z) can be represented as the product of the error signal E(z) and the transfer function 1 / A(z):

$$S(z) = \frac{E(z)}{A(z)}$$

The transfer function 1/A(z) represents an all-pole digital filter. The spectrum of the error signal E(z) will have a different structure depending on whether the sound it comes from is voiced or unvoiced. Voiced sounds are produced by vibrations of the vocal cords. Their spectrum is periodic with some fundamental frequency (which corresponds to the pitch).

These coefficients were averaged across the whole signal to give a mean coefficient vector representing the utterance. Thus, a p sized vector was used for training and testing. The value of p was based on tests given speed vs. accuracy [7]. A p value of around 10 was observed to be accurate and computationally feasible.

2.5. Distance Computation

Pattern Matching Technique is used to compare speech pattern in order to determine their similarity. In determining their similarities, speech signal is represented by the time sequence of spectral vectors. Thus we define a test pattern, T, as the concatenation of spectral frames over the duration of the speech, such that

$$T = \{t1, t2, t3,..., ti\},$$

where each ti is the spectral vector of the input speech at time i, and I is the total number of frames of speech. In a similar manner we define a set of reference patterns, {R1, R2,..., RV} where each reference pattern, Rj, is also a sequence of spectral frames, such that

$$Rj = \{rj1, rj2,...rjj\}.$$

The goal of the pattern-comparison stage is to determine the distance of T to each of the Rj, $1 < j < V$, in order to identify the reference pattern that has the minimum distance, and to associate the spoken input with this pattern [18].

The Euclidean Distance classifier uses a Euclidean distance equation to find the distance between two feature vectors. If A = (x1, x2) and B = (y1, y2) are two 2-dimensional vectors, then the distance between A and B can be defined as the square root of the sum of the squares of their differences:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

2.6. Decision Rule

Maximum Likelihood (ML) method, Baye's Theorem, is used to help decide which among the result most likely is the speaker. In this stage, average distances between the input and the codebook will be converted to percentage and is processed using ML formula. Figure 13 shows how the researchers applied the Baye's Algorithm (Maximum Likelihood Method) in this study.

Using likelihood functions, Bayes theorem is written

$$Pr(\lambda_i \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \lambda_i) Pr(\lambda_i)}{\sum_{h=1}^{n} f(\mathbf{x} \mid \lambda_h) Pr(\lambda_h)}$$

where f(x|li) is the probability density function or likelihood that a vector x is observed in model li, f(x) is the unconditional probability density function for all speaker models, and Pr(li) is the a priori probability of speaker si being the unknown speaker.

3. Testing and Evaluation

3.1. Testing

Fifty different speakers, each having six text-independent voice samples are used to perform the experiments. Five of the six samples served as training sets and is stored in the codebook. The remaining one voice sample is used as a test set to perform identification of speaker. The recordings were all done inside the Information and Resource Center of the College of Computer Studies in order to have a consistent recording. In that same room, room noise was also recorded so that the silent threshold can be set.

Because of the huge amount of data involved during identification, and so as not to jeopardize response time to generate result, the 50 different speakers were grouped into 5 by alphabetical order. Each group is composed of 10 members mixed with male and female speakers forming 5 groups of codebooks. Table 1 shows the list of names of speaker and the group used in the experiments.

Speakers				
Group 1	Group 2	Group 3	Group 4	Group 5
1. Aimee	11. Darcy	21. Gian	31. Jumalyn	41. Paulo
2. Airus	12. Dave	22. Harmz	32. Kat2x	42. Recille
3. Andrew	13. Dj	23. Ian	33. Kath	43. Remuel
4. Angel	14. Dondi	24. Jake	34. Kenneth	44. Ryan
5. Angelne	15. Elaine	25. Jam	35. Kevin	45. Rygl
6. Ana	16. Eman	26. Janine	36. Lenard	46. Shie
7. Athena	17. Feb	27. Jeca	37. Lloyd	47. Silvin
8. Chedet	18. Gail	28. Joann	38. Lurenz	48. Stella
9. Xristian	19. Gene	29. Johann	39. Luzbee	49. Tedrick
10. Cindy	20. Ger	30. Joshua	40. Merah	50. Vernon

Table 1: Speakers List and groupings

All the speakers were tested using an LPC coefficient value of 10 and 20. The testing also varies the number of voice samples in the codebook. One test includes one voice sample for each speaker stored in the codebook, and another includes five voice samples stored.

3.2. Testing Results

Figure 1 illustrates the speaker recognition performance of the system when it was tested on the five groups having one and five voice samples. The graph shows the recognition percentage of the system based upon two different LPC coefficient values and the number of voice samples each of the speaker has in the codebook. The graphs show that the application was able to deliver the right speaker within the range of 30 – 40%. It also shows further that speakers having five voice samples stored in the codebook with LPC 20 resulted to the highest percentage of recognition.

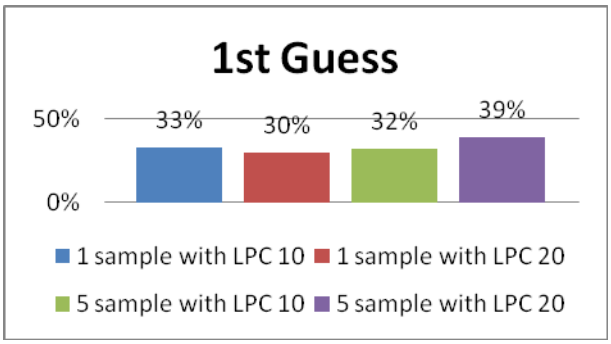

Figure 1: Closest Speaker Data Comparison

The figure below (Figure 2) is another test we performed to the system wherein the system, aside from generating a first guess, also generates a second speculation. The graph shows the recognition rate for the second guess test with LPC coefficients of 10 and 20, and with varying number of voice samples in the codebook. The second most likely speaker test result generates around 15-18% recognition.

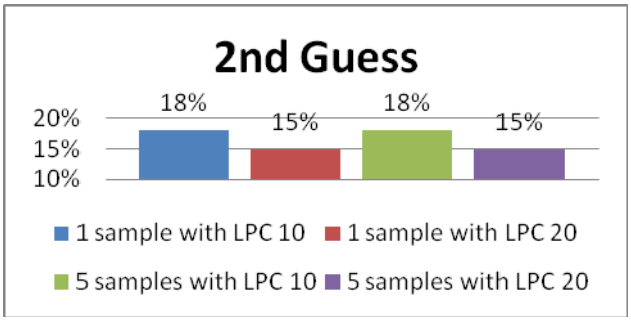

Figure 2: Second Closest Speaker Data Comparison

4. Conclusion

This study was able to yield an application able to recognize a speaker through his/her voice using pattern-matching approach. But, as evident in the testing results, the percentage of recognition is quite low. The 1st guess exceptionally surpasses the 2nd guess in terms of recognition using 5 samples, regardless the LPC used. LPC 20 though exhibited a slightly higher recognition on average but it's not true for all groups.

In this light, the researchers concludes that although it is possible to apply the pattern-matching approach to speaker recognition, this approach yields low percentage of recognition especially if the speech is not properly pre-processed. One big problem for this kind of research is the process of eliminating the noise. There is no known method to effectively do this but there is a theory how. This theory was applied in this research but the wave output after pre-process and feature extraction still displays the presence of noise.

Noise is a big hindrance for this kind of study especially that the approach chosen is pattern-matching on text-independent speech. This method uses direct comparison so the presence of noise compared against the speech yields wrong comparison which results to wrong recognition.

5. Recommendation

Based on the results, it is recommended that the algorithms used for pre-processing stage need further studies in order to totally eliminate the presence of noise. Further studies specifically on noise removal and voice filtering are highly recommended since the researchers find this area quite significant for a study like this.

It could also be on an entirely separate study the analysis of other existing algorithms to cater to text-independent speaker recognition. Although artificial intelligence is a promising field for this study, it would be better to concentrate on pattern-matching first so as to understand speech

6. References

[1] Abelgas, Minette G. [et al.], "Speech recognition using joint time frequency analysis," De La Salle University, 2002.

[2] Aldhaheri, R. and Al-Saadi, F., "Robust Text-independent Speaker Recognition with Short Utterance in Noisy Environment Using SVD as a Matching Measure," Saudi Arabia (11 February 2004).

[3] Ariyaeeinia, Sotudeh and Bailey (2004). "User Voice Identification," Amadeus Virtual Research Center.

[4] Baquiran, Eric P. [et al.], "The Applicability of linear predictive coding and vector quantization to Filipino speech recognition," University of the Philippines Diliman, 2000.

[5] Benharouga, Jalal [et. al.], "Methods for Speech Recognition," retrieved from http://www.owlnet.rice.edu/~elec532/PROJECTS98/speech/ on July 20, 2009.

[6] Campos, Diane Kristine M. [et al.], "Filipino-English bilingual speech recognition system," University of the Philippines Diliman, 2004.

[7] Clément, Ian [et. al.] (2003), "Modular Audio Recognition Framework and Text-Independent Speaker Identification, v. 0.2.0," Canada.

[8] Cook, Stephen (2002). Speech Recognition HowTo.

[9] Frederic Bimbot [et al.], "A Tutorial on Text-Independent Speaker Verification," EURASIP Journal on Applied Signal Processing 2004:4, 430–451.

[10] Guevara, Rowena Cristina [et al], "Development of a Filipino Speech Corpus," Digital Signal Processing Laboratory, University of the Philippines, 2002.

[11] Gupta, Harsh [et. a.], "Field Evaluation of Text-Dependent Speaker Recognition in an Access Control Application," Speech and Image Processing Unit, Department of Computer Science, University of Joensuu, Joensuu, Finland, 2005.

[12] Hagai Aronowitz, David Burshtein, "Efficient Speaker Identification and Retrieval," Israel (2005).

[13] Kofi Boakye and Barbara Peskin, "Text-Constrained Speaker Recognition on a Text-Independent Task," International Computer Science Institute, Berkeley, CA USA (2004).

[14] Montenegro, Chuchi S., "Probabilistic Speech Recognition for Tagalog Lecture Video (PSRTLV)" College of Computer Studies, Silliman University, Dumaguete City, 2008.

[15] Navarro, Rolando D. Jr., "Recognition of Tagalog Alphabets Using The Hidden Markov Model" Presented at the 10th National Convention on Statistics (NCS) at EDSA Shangri-La Hotel 1-2 October 2007.

[16] NCH Software. Audio File Formats, Retrieved May 25, 2009 from http://www.nch.com.au/acm/formats.html

[17] Reynolds, Douglas A. & Heck, Larry P. "Automatic Speaker Recognition: Recent Progress, Current Applications, and Future Trends" 19 February 2000 Presented at the AAAS 2000 Meeting Humans, Computers and Speech Symposium 19 February 2000.

[18] Rabiner, Lawrence & Juang, Biing-Hwang (1993). Fundamentals of Speech Recognitio (New Jersey: Prentice-Hall Inc., 1993).

[19] "Speaker Recognition," 7 June 2005, National Science and Technology Council.

[20] Sumantray, Ronak Kumar, "Automated Attendance System: With a New Approach to Voice Identification Method," College of Engineering and Technology, Bhubaneswar (2005).