World Applied Sciences Journal 31 (10): 1767-1771, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.31.10.468

Speech Fingerprint to Identify Isolated Word-Person

Khaled Matrouk, Abdullah Al-Hasanat, Haitham Alasha'ary,
Prof Ziad Al-Qadi, Prof Hasan Al-Shalabi

Department of Computer Engineering, Faculty of Engineering,
Al-Hussien Bin Talal University, P.O. Box (20), Ma'an, Jordan

Abstract: Voice recognition is the process of taking the spoken word as an input to a computer program. This process is important to virtual reality because it provides a fairly natural and intuitive way of controlling the simulation while allowing the user's hands to remain free. This paper will delve into the uses of voice recognition in the field of virtual reality, examine how voice recognition is accomplished and list the academic disciplines that are central to the understanding and advancement of voice recognition technology. In order to recognize the spoken word and the person who spoke the word some estimating parameters were chosen and calculated to form voice fingerprint. The calculated values of voice fingerprint parameters then were passed to artificial neural networks, it was shown that using parameters values achieve high recognition rate, thus identifying the spoken word and the speaker.

Key words: Sigma · Mu · Peak factor · Dynamic range · Power spectral density · Zero crossing rate · Artificial neural network · Voice recognition

INTRODUCTION

Voice recognition is "the technology by which sounds, words or phrases spoken by humans are converted into electrical signals and these signals are transformed into coding patterns to which meaning has been assigned[1 and 2]. While the concept could more generally be called "sound recognition", we focus here on the human voice because we most often and most naturally use our voices to communicate our ideas to others in our immediate surroundings. In the context of a virtual environment, the user would presumably gain the greatest feeling of immersion, or being part of the simulation, if they could use their most common form of communication, the voice. The difficulty in using voice as an input to a computer simulation lies in the fundamental differences between human speech and the more traditional forms of computer input. While computer programs are commonly designed to produce a precise and well-defined response upon receiving the proper (and equally precise) input, the human voice and spoken words are anything but precise. Each human voice is different and identical words can have different meanings if spoken with different inflections or in different contexts. Several approaches have been tried, with varying degrees of success, to overcome these difficulties [1].

We can classify speech recognition tasks and systems along a set of dimensions that produce various tradeoffs in applicability and robustness [3,4].

Isolated Word Versus Continuous Speech: Some speech systems only need identify single words at a time (e.g., speaking a number to route a phone call to a company to the appropriate person), while others must recognize sequences of words at a time. The isolated word systems are, not surprisingly, easier to construct and can be quite robust as they have a complete set of patterns for the possible inputs. Continuous word systems cannot have complete representations of all possible inputs, but must assemble patterns of smaller speech events (e.g., words) into larger sequences (e.g., sentences).

Corresponding Author: Haitham Alasha'ary, Department of Computer Engineering, Faculty of Engineering, Al-Hussien Bin Talal University, P.O. Box (20), Ma'an, Jordan, Tel: +962-3-2179000 (ext. 7523), Fax: +962-3-2179050, Cell: +962-776-247865.

Speaker Dependent Versus Speaker Independent Systems: A speaker dependent system is a system where the speech patterns are constructed (or adapted) to a single speaker. Speaker independent systems must handle a wide range of speakers. Speaker dependent systems are more accurate, but the training is not feasible in many applications. For instance, an automated telephone operator system must handle any person that calls in and cannot ask the person to go through a training phase before using the system. With a dictation system on your personal computer, on the other hand, it is feasible to ask the user to perform a hour or so of training in order to build a recognition model.

Small Versus Vocabulary Systems: Small vocabulary systems are typically less than 100 words (e.g., a speech interface for long distance dialing) and it is possible to get quite accurate recognition for a wide range of users. Large vocabulary systems (e.g., say 20,000 words or greater), typically need to be speaker dependent to get good accuracy (at least for systems that recognize in real time). Finally, there are mid-size systems, on the order to 1000-3000 words, which are typical sizes for current research-based spoken dialogue systems. Some applications can make every restrictive assumption possible. For instance, voice dialing on cell phones has a small vocabulary (less than 100 names), is speaker dependent (the user says every word that needs to be recognized a couple of times to train it) and isolated word. On the other extreme, there are research systems that attempt to transcribe recordings of meetings among several people. These must handle speaker independent, continuous speech, with large vocabularies. At present, the best research systems cannot achieve much better than a 50% recognition rate, even with fairly high quality recordings.

To analyze the voice signal and to create voice fingerprint which can be used to recognize the speech isolated word and the speaker person we can use the following parameters:

- Estimating the mu of the population (mu). This graphic shows an overview of all the relationships. In step 1 (in the upper left-hand corner of the graphic) you can see that the dependent variable, Sampling Distribution of the Mean(SAQ), has been modeled as a normal distribution. In step 2 we do a research project on spatial ability; this is equivalent to taking a sample of certain size, n, from this population of SAQ scores. So we've got a sample. In step 3 we calculate a statistic on the sample data. In this case we calculate the mean [5].

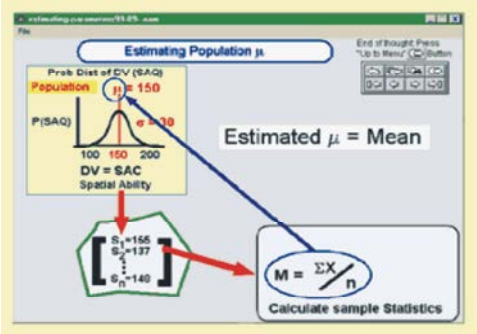

Fig. 1: Calculating mu parameter

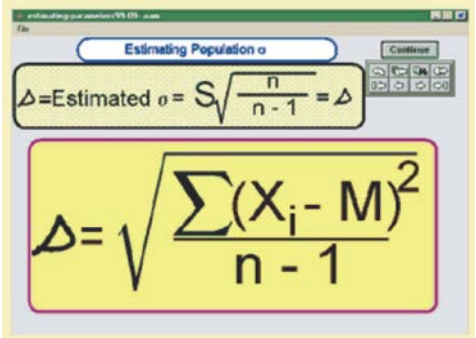

Fig. 2: Calculating sigma parameter

The estimate of the population mean, mu, is the sample mean. That's as simple as it can be. Unfortunately, it's going to be messier when we get to estimating population variance as shown in Figure (1).

- Estimating sigma. The next population parameter we want to estimate is the standard deviation, sigma. The current screen gives the formula for calculating an estimate of population standard deviation (sigma) from sample data as shown in Figure(2) [6].
- Peak Factor (Crest Factor): The crest factor of an audio signal is the dB difference between the peaks and the RMS value of the signal. The RMS (Root Mean Square) is defined as the "heating value" of the signal-the voltage that would generate the same heat as a DC (Direct Current) signal, over the same time[7]. The RMS value of a complex signal must be read with an RMS voltmeter. Alternatively, the signal can be digitally sampled and the samples summed to yield the RMS value. As such, the RMS value of a complex signal can be thought of as the "area under the curve" of a signal as viewed in a wave editor software application.
- Dynamic range, abbreviated DR or DNR is the ratio between the largest and smallest possible values of a changeable quantity, such as in signals like sound

and light. It is measured as a ratio, or as a base-10 (decibel) or base-2 (doublings, bits or stops) logarithmic value [8].

- Power spectral density(PSD). For continued signals that describe, for example, stationary physical processes, it makes more sense to define a *power spectral density* (PSD), which describes how the power of a signal or time series is distributed over the different frequencies, as in the simple example given previously. Here, power can be the actual physical power, or more often, for convenience with abstract signals, can be defined as the squared value of the signal [9,10].
- Zero-crossing rate. The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds [11].

MATERIALS AND METHODS

To obtain a speech fingerprint the following methods and tools were used:

- Personnel computer
- Matlab package.
- Matlab programs and function to calculate the above mentioned voice parameters.
- Matlab artificial neural network (ANN) for word and person identification.

Experimental Part: The experiment was divided into three parts:

First Part: Calculating speech parameters

A matlab code and functions were built and executed to find the fingerprint for each word-person by calculating various values of speech parameters some of these values are shown in Table 1 through Table 6.

Second Part: Creating and Training ANN: The experimental data obtained in the first part were used to form the input matrix of ANN. This matrix contained 6 rows and 900 columns, each column contained the parameter values of the spoken word-person (5 persons and 6 words). This input matrix was used to train ANN with the following specifications: one input layer with 6 neurons (one for each parameter value), one hidden layer with 36 neurons, one output layer with two neurons (the first one to identify the person and the second one to identify the word), tansig activation function and trainlm training function.

Third Part: Testing ANN: The created and trained ANN in the previous part was tested using sample data for different word-person. One hundred fingerprint for each word-person were used to test ANN and the recognition ratio was calculated. The obtained results are shown in Tables 7 and 8.

The above tested data were used also to calculate the recognition ratio of identifying word-person and the worst value obtained was equal 96.2.

Table 1: Parameters for isolated word yes for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	PSD	Zero Crossing rate
Esraa	0.18212	-0.0026199	14.7924	84.201	2.0323e-006	700
Alaa	0.21686	-0.0018049	13.2764	90.2216	2.8809e-006	705
Ayyam	0.22233	-0.00098613	13.0604	84.201	3.0277e-006	619
Heyam	0.25442	0.00053341	11.8894	80.6792	3.9647e-006	660
Mohammad	0.13602	-0.0012675	17.3275	90.2216	1.1335e-006	558

Table 2: Parameters for isolated word no for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	Power spectral density PSD	Zero Crossing rate
Esraa	0.20339	-0.0001826	13.8337	84.201	2.5338e-006	693
Alaa	0.21477	0.0033057	13.3599	90.2216	2.8265e-006	775
Ayyam	0.21846	0.00021444	13.2129	84.201	2.9231e-006	555
Heyam	0.27033	0.00040072	11.3624	80.6792	4.4762e-006	758
Mohammad	0.19006	-0.0016707	14.422	90.2216	2.2130e-006	640

Table 3: Parameters for isolated word up for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	PSD	Zero Crossing rate
Esraa	0.12378	-0.00041674	18.1472	90.2216	9.3849e-007	706
Alaa	0.19224	0.0019817	14.3228	90.2216	2.2642e-006	583
Ayyam	0.11292	-0.00023592	18.9446	90.2216	7.8106e-007	652
Heyam	0.20545	0.0011952	13.7462	84.201	2.5855e-006	612
Mohammad	0.12595	0.00065475	17.9965	90.2216	9.7164e-007	634

Table 4: Parameters for isolated word down for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	PSD	Zero Crossing rate
Esraa	0.17407	0.002138	15.1854	90.2216	1.8564e-006	648
Alaa	0.1933	0.00092033	14.2753	90.2216	2.2888e-006	833
Ayyam	0.15728	0.0006252	16.0668	90.2216	1.5152e-006	760
Heyam	0.22434	-0.0017447	12.9819	84.201	3.0830e-006	762
Mohammad	0.13476	0.0028708	17.4069	90.2216	1.1134e-006	778

Table 5: Parameters for isolated word left for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	PSD	Zero Crossing rate
Esraa	0.20498	-0.0024832	13.7653	84.201	2.5744e-006	575
Alaa	0.19427	-0.00032522	14.232	90.2216	2.3118e-006	652
Ayyam	0.20868	4.3797e-005	13.6105	84.201	2.6674e-006	584
Heyam	0.23733	-0.0062569	12.4902	84.201	3.4547e-006	646
Mohammad	0.12563	-0.00053263	18.0186	90.2216	9.6668e-007	556

Table 6: Parameters for isolated word right for different persons

Name	sigma	mu	Peak (crest) factor Q (dB)	Dynamic range D (dB)	PSD	Zero Crossing rate
Esraa	0.16287	0.0059857	15.7577	90.2216	1.6291e-006	767
Alaa	0.19439	6.1914e-005	14.227	90.2216	3.0248e-006	931
Ayyam	0.22212	-0.0048979	13.0666	84.201	3.0248e-006	690
Heyam	0.24064	-0.0016376	12.3725	84.201	3.5474e-006	789
Mohammad	0.12773	-0.00010019	17.8744	90.2216	9.9932e-007	737

Table 7: Recognition ratio for each word

Word	Testing samples	Correct recognition	Incorrect recognition	Recognition ratio(%)
Yes	500	500	0	100
No	500	500	0	100
Up	500	494	6	98.8
Down	500	488	12	97.6
Left	500	500	0	100
Right	500	500	0	100

Table 8: Recognition ratio for each person

Person	Testing samples	Correct recognition	Incorrect recognition	Recognition ratio(%)
Esraa	600	600	0	100
Alaa	600	600	0	100
Ayyam	600	587	13	97.8
Heyam	600	592	8	98.6
Mohammad	600	585	15	97.5

CONCLUSIONS

From the obtained results we can conclude the following:

- The chosen parameters are suitable to create a speech fingerprint.
- The created speech fingerprint can be used to recognize isolated word, person and word-person with high recognition ratio.

REFERENCES

1. "Speaker Independent Connected Speech Recognition-Fifth Generation Computer Corporation". Fifthgen.com. Retrieved 2013-06-15.
2. Jump up "British English definition of voice recognition". Macmillan Publishers Limited. Retrieved February 21, 2012.
3. Sadaoki Furui, 2005. 50 years of Progress in speech and Speaker Recognition Research, ECTI Transactions on Computer and Information echnology, Vol. 1. No. 2 November 2005.
4. Juang, B.H. and R. Lawrence, 2004. Rabiner, Automatic Speech Recognition-A Brief History of the Technology Development, October 8, 2004.
5. Richard, M. Stern, Fu-Hua Liu, Yoshiaki Ahshima, Thomas M. Sullivan and Alejandro Acero, Multiple Approaches To Robust Speech Recognition.
6. Wolf, R., F. Ellinger, R. Eickhoff, Massimiliano Laddomada, Oliver Hoffmann (14 July 2011). Periklis Chatzimisios, ed. Mobile Lightweight Wireless Systems: Second International ICST Conference, Mobilight 2010, May 10-12, 2010, Barcelona, Spain, Revised Selected Papers. Springer. pp: 164. ISBN 978-3-642-16643-3. Retrieved 13 December 2012.
7. Bin, Wu, Jianwen Zhu and Farid Najm, 2006. Dynamic Range Estimation. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, pp: 1618-1636.
8. Bin Wu, Jianwen Zhu and Farid Najm, 2004. An analytical approach for dynamic range estimation In ACM/IEEE 41st Design Automation Conference (DAC-04), San Diego, Calif., June 7-11, 2004.
9. Bin Wu, Jianwen Zhu and Farid Najm, 2004. Dynamic range estimation for nonlinear systems. In IEEE/ACM International Conference on Computer-Aided Design (ICCAD-04), San Jose, Calif., November 7-11, 2004.
10. Scott Millers and Donald Childers, 2012. Probability and random processes. Academic Press, pp: 370-5.
11. Gouyon, F., F. Pachet and O. Delerue, 2000. Classifying percussive sounds: a matter of zero-crossing rate?, in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, December 7-9, 2000. Accessed 26th April 2011.