CS570 - Artificial Intelligence Research Paper

Chihsiang Wang

# Voice recognition

## By Chihsiang Wang

**Introduction**

Voice Recognition has become a hot commodity in modern society as of late. It is common to see VR applications everywhere, from the PDA to the smart phone. It has wormed its way into our lives. The use of VR in smart phones is a most commons case in today. In 2011, Apple Inc. launched the Voice Recognition secretary “Siri” as the newest smart product. It has revived the human desire to talk to machine indirectly again. The following year, Google released the voice recognition search engine “Google Search” in the newest smart phone’s OS “Jelly Ben”. It applies to ANN and reduces 25% of recognition faults **[9]**. But the VR system continues to have several difficulties it needs to overcome. For example, the VR system is more effective in an ideal and controlled situation. In real life, however, variables such as languages, pronunciations, and noises occur; even a recording machine can cause different speech signals. Through hard work and creative thinking, such obstacles can be overcome. Technology companies can see the good economic benefit and value of VR technique, and thus, invest in its research.

**Abstract**

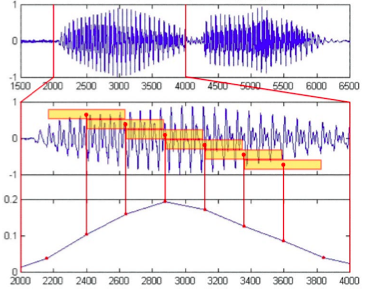
The voice is one of most basic tools for humans to exchange information; it’s fast and effective. A system that can analyze, understand and convert voice commands into tangible actions is the ideal goal. Although computer technology has developed quickly, and computer processors’ speed has increased, the artificial intelligence (A.I.) of computers still lack the ability to analyze semantics. In recent research of the voice recognition (VR) system, most program designers focus on converting the source of the sounds to be specific words or models they exclude analysis of the whole sentence because the semantics is too complex to result in clear definitions **[3]**.

The voice recognition system doesn’t require specific programming languages, and most systems are built with similar algorithms **[5]**. Dynamic Time Warping (DTW) was the mainstream identifying method in early researching. It considers the difference of the talking speed, and makes appropriate accommodations. Artificial Neural Networks (ANNs) can often identify voice input with sufficient training. Hidden Markov Model (HMM) applies statistics to describe the symbol of the voice, and analyzes it for recognition. There are some classifications to define VR systems. One of the most common ways to classify VR systems is the categorization of speaker-dependent and speaker-independent systems. The speaker-dependent system is designed for particular users. The system customizes the learning system to improve recognition accuracy, but it cannot be guaranteed for other users. The speaker-independent system is designed for normal users. With this system, there is more room for error, because the system requires a huge collection of samples. To overcome this problem, the speaker-independent system uses speaker adaptation functions. In other words, the first time the VR system is used, it requires users to record a couple sentences to build the database and create the adjusting individual parameter **[2]**. When users continue to operate the system, it loads more parameters to increase recognition accuracy. The VR system still has several difficulties it needs to overcome: different languages, pronunciations, background noise, continuous sentence, etc. The ideal VR system lets the human user talk to machine as a real human **[2]** —an ideal similar to that of artificial intelligence.

**Window Function**

The human vocal cords change in shape and position while talking and make different sounds. As a result, the vocal system can be sorted as a “time-varying system” **[2]**. The sounds that are emitted are non-stationary signals, or signals that change over time. On the other hand, the vocal chords may not change in shape and location in a small period of time—about 20 to 30 ms **[2]**. Assuming this case, the vocal system is a “time-invariant system” **[2]** for a given time increment; in the other words, a vocal system can be classified to be a “piece-wise stationary signal” **[2]**.

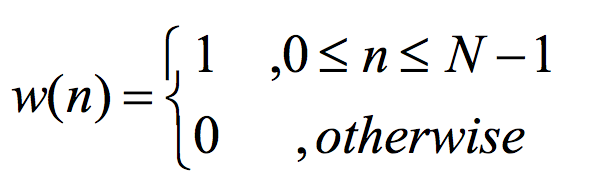
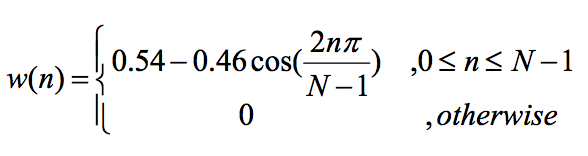
For a discrete voice signal, x(n), select a sound wave, analyze it, and obtain the “speech features”. The selected portion of speech is called a “sound frame”**[5]**. A whole sound signal can be split into several continuous speech frames. When this concept is applied to the human vocal system, it is usual to receive approximately 20 to 30 ms per frame**[2]**. Each frame overlaps into the next by about one-third to one-half frame lengths. The overlaps can help to observe changes in between the sound frames.

**

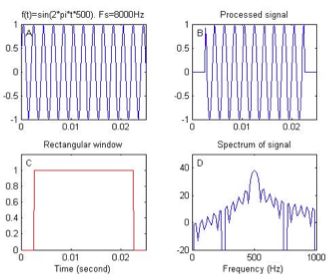
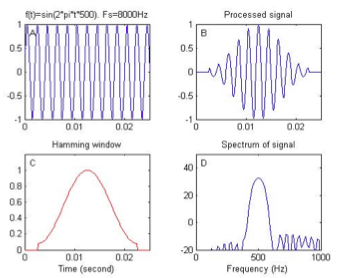
**Example of sound frame processed [6]**

The sound frames lengths affect the accuracy of the voice feature parameters. Long frames have more sample points to analyze. If the frames, however, are too long, it will be difficult to observe any changes. The two most common kinds of frames that are applied to voice recognition are the Rectangular window and the Hamming window. Assuming the length of window is ***N***, the mathematic formulas are listed below:

The Rectangular window is the most frequently used method. When a window is applied to a speech signal, only the part in window will be reserved and analyzed, otherwise the speech signal is set as zero. The Rectangular window, however, cuts the speech signal into discontinuous frames causing a reduction in analysis ability. Different from the Rectangular window, the Hamming window can reduce the discontinuity by fading out sides of each frame when applied to the voice signal.

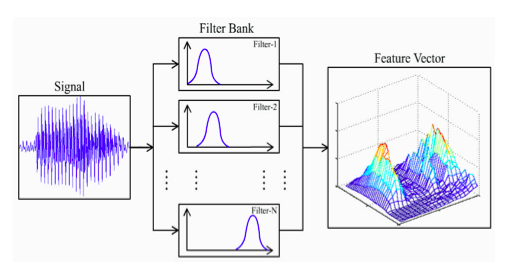
 

Rectangular window Hamming window

 **[5]**

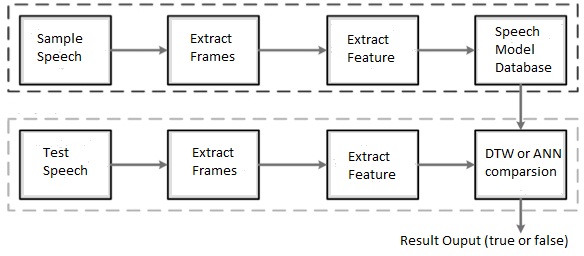
**Speech Feature Extraction**

Each person has his or her own individual vocal system. Internal and external variants can make different sounds, even if the sounds come from the same person. As a result, comparing speech signal wave figures is useless. Voice recognition systems, therefore, extract “feature vectors”. Feature vectors are reflected sounds from each speech frame. It can replace the original figure to obtain a better efficacy to analyze. It is very important to choose a speech feature in every kind of voice recognition systems. An important feature vector in voice recognition systems is the “Linear-Frequency Cepstral Coefficient” (LFCC)**[6]**. LFCC can help to attain better accuracy in identification for voice recognition. To extract a speech feature, the most direct method is to send a vocal signal into a “digital band-pass filter bank” **[6]** so that it is divided into several partitions for conversion. The output is the feature vector.



**Digital band-pass filter bank [6]**

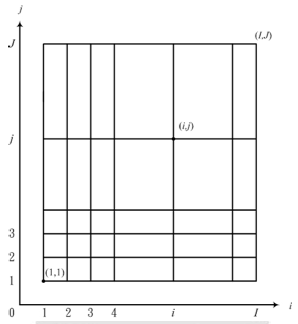
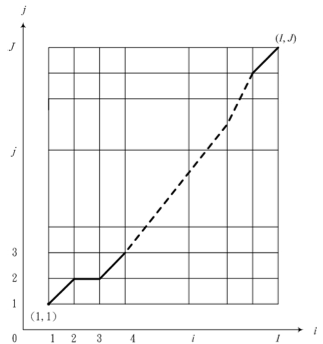
With the existence of speech feature vectors, it has become easier to do comparisons. Today, the most common methods for voice recognition are Dynamic Time Warping and Artificial Anural Network. The complete voice recognition process figure is shown below.



**Process of voice recognition [5]**

**Dynamic Time Warping**

To compare two speech signals, the speech signals’ feature needs to be extracted. Using the Hamming window, the speech signal is split into an array of continuous speech frames; which, are then extracted for comparison. In the figures below, vectors on the i-axis are the input (test) speech—which are all I frames—while vectors on the j-axis are the reference (sample) speech—which are entirely J frames. The (i,j) shows the i(th) and j(th) frames to compare **[6]**.

**Comparison figure** **Best path comparison [6]**

Unfortunately, the test speech and sample speech have different lengths so they cannot be compared. Applying a dynamic programming method to adjust the i-axis, the two frames now have more similar lengths This method is called dynamic time warping; it finds the best path from (l,l) to (I,J) **[6]**. Each point on the path explains the speech features at their corresponding moments in time. If the test speech signal transitions one grid-length on i-axis but not the j-axis, the sample speech has quicker changes. Inversely, if the test speech signal moves one grid-length on j-axis but not i-axis, the sample speech changes at a slower rate. To define a variable , the distance between the test speech and sample speech;  is the weight of the transition from  to . The entire distance of path ***D*** can listed as . A lower distance shows a smaller difference **[1]**.

**Artificial Anural Network**

Artificial Anural Network (ANN) is an information exchange system, which imitates the human anural network. ANN, in general, can be either a supervised learning network or an unsupervised learning network. In early voice recognition research, DTW was used. It is being progressively replaced by the Back Propagation Algorithm (BPNN) system, a supervised learning network. Both BPNN and DTW have similar success rates and recognition accuracy of speech comparisons; but BPNN has better recognition speed and requires little memory resources. In a recent research study, where the voice recognition setting was a controlled variable, BPNN’s recognition speed is 120 times faster than DTW. Also, BPNN’s use of memory resources is only 24% to that of DTW **[5]**.

A BPNN system uses a gradient descent method to train the system to minimize the difference between the output and the target. There are three layers within the system: the input layer, the hidden layer(s), and the output layer. The output layer has two nodes in the VR system – true and false.

The process of a complete voice recognition system must first receive a test speech. A Hamming window is then applied to the test speech in order to extract specific speech frames. After a pre-emphasis work and reduction of external noisy signals, a pure speech signal is produced to extract frame feature vectors. The BPNN system resulted frame feature vectors are used as training parameters for the ANN model. The ANN model receives adequate training from the parameters to build a database of sample speeches. Once the trained system is fully built, test speech feature vectors can be inputted for a voice comparison. The training procedure of the BPNN system has four steps:

|  |  |
| --- | --- |
| P | The i(th) input |
| A | The output of weight |
| O | The output value |
| F | The function used to convert the anural nodes |
| W | The weight connects each layer. |

1. Set up the network structure, define arguments for training times, and give the default input weights. In a general case, the default input weight can be any value, but it usually is in between 0.5 ~ -0.5 or 1 ~ -1 **[5]**.
2. Calculate each node’s output by the feed forward method:

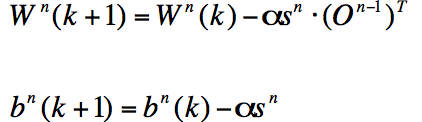


Note that  (input nodes) and  (output nodes).

1. Use the feed backward method by Back Propagation Algorithm to calculate the sensitive parameters (errors):



1. With the O and S values obtained from Steps 1 and 2, use the Steepest Descent Algorithm to adjust each node’s weights in each layer until convergence. Every time a training sample is received, the system will update and adjust the weights in each nodes.



Repeat step 2 through 4 until the loop reaches the statement to stop training. An ANN system completion depends on the length of the training time. There are two methods to set up training time: giving a specific time frame for the training, or set a statement to stop the training of the system. Both methods, however, have their own risks of errors. In the first method, if the users do not give an adequate amount of time for training, the ANN system cannot receive enough samples to adjust the nodes into the correct weights. In other words, the ANN system cannot be “fully trained” to efficiently attain recognition accuracy. In the second method, if an irrelevant argument is given, it is easy to enter the infinity loop and never stop training. To overcome this situation, more than one relevant argument is given **[5]**.

**Conclusion**

Released in 1986, the BPNN helps the ANN to overcome many difficulties. The ANN, however, still needs to learn more information and have a more efficient calculation of time. It is still far away from the ideal system—voice recognition in real time. Contrary to the ANN, fuzzy logic is another method to imitate the human neural network. It has been used in many fields for its advantage of a fast learning rate; but it has lower fault toleration.

Today, Fuzzy Logic has been used in automatic control, pattern recognition, and broad-based decision analysis. It imitates the human’s logical thinking in spirit, so it does not need a high precision math algorithm like the ANN. It uses a person’s knowledge as the control rules. As a result, it is easy to adjust the rules and attain the expected results. These characteristics reduce huge difficulties in designing the control system. For example, a washer installed with a fuzzy chip is very common nowadays. These washers can automatically detect how heavy the laundry load is and how much detergent should be used. All users need to do is to select the full-auto button, and then the logic chip will do the rest in very short time like an expert operator. Fuzzy logic, however, requires a large collection of information to build rules. It cannot train or learn by itself—the designers need to define the rules. This is different from the ANN. ANN and Fuzzy Logic is both useful and reliable technology in the present day. ANN can be trained to produce good results when it has an adequate knowledge base. Fuzzy Logic has an efficient ability for technology control. The NeuroFuzzy is the hybrid technology of ANN’s and Fuzzy Logic’s advantages **[10]**. The use NeuroFuzzy, consisting of high recognition accuracy, high noise tolerance, and short reflection time, to create a voice recognition system is a valuable field of study for future research.

**Reference**

[1] C. Kim and K. Seo, “Robust DTW-based Recognition Algorithm for Hand-held Consumer Devices”,IEEE Transactions on Consumer 706 Electronics, Vol. 51, pp.699-709, 2005.

[2] H. Sakoe and S. Chiba, “Dynamic Programming Optimization for Spoken Word Recognition”, IEEE Transactions on ASSP, Vol.26, pp 43-49, 1978.

[3] G.D. Wu and C.T. Lin, “A Recurrent Neural Fuzzy Network for Word Boundary Detection in Variable Noise-Level Environments,” IEEE Trans on System, Man, and Cybernetics, Vol. 31, No. 1, pp 84-97, Feb. 2001.

[4] A. Hussain, S.A. Samad and L.B. Fah, “Endpoint Detection of Speech Signal using Neural Network,” IEEE Trans on ASSP, pp 271-274, 2000.

[5] X.L. Chen, “The Speech Recognition System using Neural Networks”, Chinhwa national university. 2011.

[6] C.C. Zhou, Improved DTW-based Speech Recognition System with Its FPGA Implementation and Analysis”, Taiwan national university. 2009.

[7] Talking to Machines. (2012). Communications of the ACM, 55(4), 14-16. doi:10.1145/2133806.2133812

[8] NO SURPRISE, SIRI SATISFIES. (2012). EventDV, 2012(5), 10.

[9] Rodriguez, A. (2012). FAQ: Google Now and Android Jelly Bean Voice Recognition. PC World, 30(9), 18.

[10] Wired Tech Research (2012). Retrieved from

<http://wired.tw/2013/04/23/neutralnetworksandroid/index.html>

[11] Taiwan National Academic For Education Research (2012). Retrieved from

http://terms.naer.edu.tw/detail/1679037/