CS570 - Artificial Intelligence Research Paper

Chihsiang Wang

# Voice recognition

## By Chihsiang Wang

**Summary**

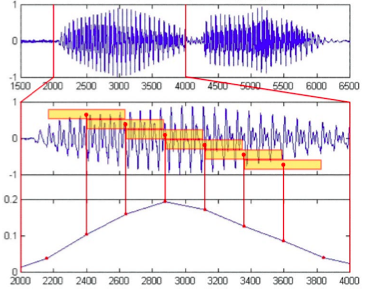
Voice is one of most basic abilities for human to exchange information; it’s fast and effective for human to get a better communication quality. A system that can analyze, understand and convert voice commands in to actual actions is an ideal target. However, although computer technology has developed quickly, and computer processors’ speed have increased, the A.I. (artificial intelligence) of computers still lack the ability to analyze semantics. In the recent research of voice recognition (VR) system, most of program designers focus to convert resource of sound to be a particular words or models, but not analyze the significance of whole sentence, because the semantics is too complex to give a clearly definition.

The voice recognition system doesn’t require for specific programming languages, and most of systems build on similar algorithms. Dynamic Time Warping (DTW) was the mainstream identify method in the early researches, it considers the difference of the talking speed, and make an appropriate compensation for the system. Artificial Neural Networks (ANNs) can often identify voice input with sufficient training. Hidden Markov Model (HMM) applies statistics to describe the symbol of voice, and analyze it for recognition. There are some classifications to define VR systems; one of the most commons ways classifies VR system as speaker-dependent and speaker-independent system. Speaker-dependent system is designed for particular users, the system customizes the learning system to improve recognition accuracy, but this system can not be guaranteed for other users. Speaker-independent system is designing for normal users, it has more difficulties to implement, because the system requires collecting huge of samples, so it is easier to cause errors than speaker-dependent system. To overcome this problem, speaker-independent used to have speaker adaptation functions, in other words, when first use the VR system, it requires users to record couple sentences to build the database, and create the adjusting parameter for each users. When users operating the system, it can load these parameters to increase recognition accuracy. Voice recognition system still have bunch of difficulties need to overcome: different language, pronunciations, background noise, continuous sentence…etc. The ideal system of VR can let human “talk to machine as real human”, and that’s same ideal as artificial intelligence.

**Window Function**

Human’s vocal organ changes its shape and position while talking, and makes different sounds, so vocal system can be sorted out as “time-varying system”; and the sound are belong to non-stationary signal, it means the signal changes with time. However, vocal organ won’t change its shape and location too much in a split time (about 20~30 ms), apply with this feature can assume that the vocal system is a “time-invariant system” in each time zone; in the other words, a vocal system can be classified to be a “piece-wise stationary signal”.

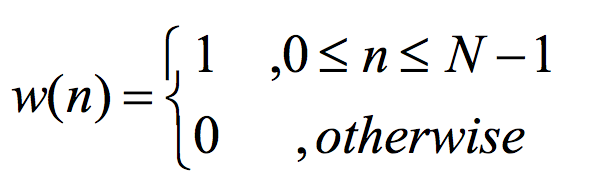
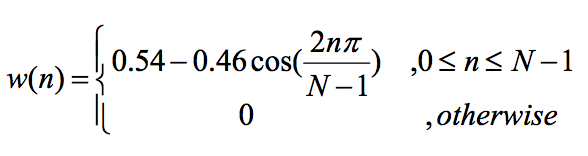
For a discrete voice signal x(n), use a window cover it and analyze it, and get the “speech features”, the part of speech is called a “sound frame”. In a whole sound signal can be split to several continuous speech frames, applies to the feature of human’s vocal system, it’s usually to set up by 20~30ms in one frame, and each frame will be overlap with next, the length of overlap usually to set up with 1/3 ~ 1/2 speech frames’ length, this can help to observe the change between each speech frames.

**

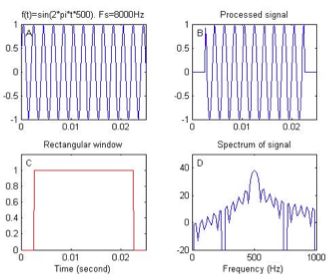
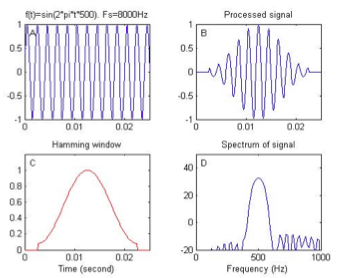
**Example of sound frame processed**

The speech frames length affect the accuracy of the voice feature parameters, longer frames get more sample points to analyze, but a too long window will be hard to observe the change inside. The most two commons windows apply to voice recognition are Rectangular window and Hamming window. Assume the length of window is ***N***, the mathematic formulas lists below:

Rectangular window is the most immediately method, when a window applies to a speech signal, only the part in window will be reserved and analyzed, otherwise will be set as 0. But rectangular window will hard cut off the frame side directly, and it will cause the speech signal discontinuously, in the other words, it reduces the ability to analyze. Different as rectangular window, Hamming window can reduce the breaks of a frame; it fades out both side of the frame while apply to the voice signal.

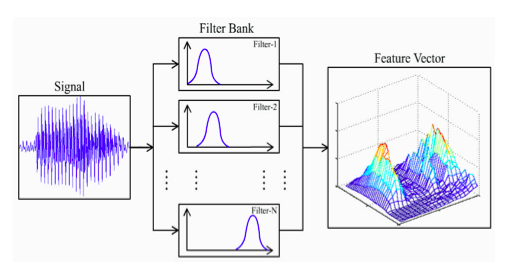
 

Rectangular window Hamming window

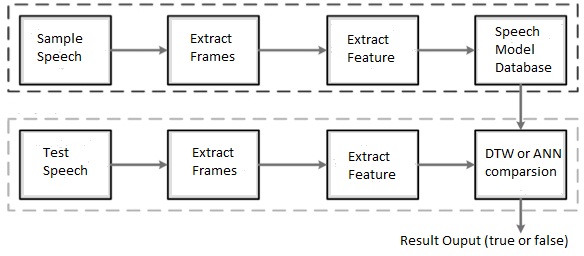
**Speech Feature Extraction**

Every persons has their own different vocal system, inner and outer effects can makes different sounds, even it comes from a same person; so it’s not worth it to compare the speech signal wave figures, in the voice recognition system, it used to extract “feature vector”. Feature vectors are reflected to the sound from each speech frames, it can replace the original figure to get a better efficacy for analyzing. It’s very important to choose a speech feature in every kinds of voice recognition system, an important feature vector applies to voice recognition system is “Linear-Frequency Cepstral Coefficient” (LFCC), it can help to get a better identify accuracy for voice recognition. To extract a speech feature, the most directly method is to send a vocal signal into a “digital band-pass filter bank” to divide the single to several partitions and convert it, the output of the filter is the feature vector.



**Digital band-pass filter bank**

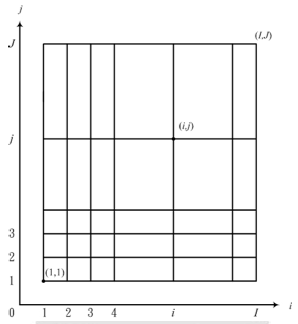
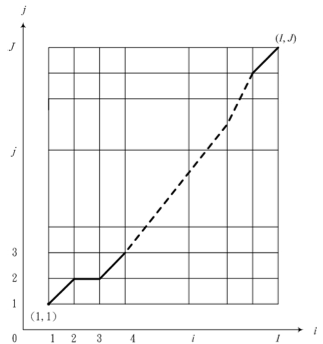
With speech feature vectors, it becomes easier to do comparison. In the present time, the most common methods to do voice recognition are Dynamic Time Warping and Artificial Anural Network. The complete voice recognition process figure shows below.



**Process of voice recognition**

**Dynamic Time Warping**

To compare two speeches, it needs to extract each speech signals’ feature vectors. So first it need to apply the hamming window and split a speech to an array of continuous speech frames, and extract each frames’ speech features for comparing. In the figure, vectors on i-axis are input (test) speech, it has totally I frames, and reference (sample) speech on the j-axis, has totally J frames, (i,j) shows the i(th) and j(th) frames to compare.

**Comparison figure** **Best path comparison**

But the test speech and sample speech has different length, it can’t be compared directly. Applying to dynamic programming method to adjust the i-axis and let two frames have more similar length, this method calls dynamic time warping, it can find a best path from (l,l) to (I,J). Each points on the path explains the moment of speech features in the same moment, if test speech signal gets one grid transition on i-axis but j-axis doesn’t, it explains the sample speech has a faster changes, on the other hands, if it moves on j-axis but not i-axis, the sample speech goes slower. To define a variable , it’s the distance between the test speech and sample speech;  is the weight of the transition from  to , the whole distance of the path ***D*** can listed as . A lower distance shows it has a lower difference.

**Artificial Anural Network**

Artificial Anural Network (ANN) is an information exchange system, which imitate human’s anural network; ANN in general can be sorted of to be supervised learning network and unsupervised learning network. In the voice recognition research, it used to use Back Propagation Algorithm (BPNN) system, and it’s belong to supervised learning network. A BPNN system in the recent research replaced DTW progressively, BPNN and DTW has similar success rate and recognition accuracy of speeches comparison, but BPNN has a much better recognition speed and it requires lower memory resources. In a recent research experiment shows that in the same setting to do voice recognition, BPNN’s recognition speed is 120 times faster than DTW, and BPNN only requires 24% of memory resources than DTW. It also notices that BPNN has a better fault tolerant ability for the input speech in the same experiment.

A BPNN system uses gradient descent method to train the system to minimize the difference between the output and the target, it has three layers in the system: input layer, hidden layer(s), output layer, in the output layer has two nodes in the VR system – true and false. The process of a complete voice recognition system will get the test speech, applies to the hamming window and extract particular parts of voice frames, after a pre-emphasis work and reduce outer noisy signals, it gets the pure speech signal for extracting the frame feature vectors. In the BPNN system, frame feature vectors are using to for the training parameters; an ANN model gets enough training with these parameters, will build the database of the sample speeches. After build a full trained system, it just needs to input the test speech feature vectors to do the voice comparison. The training procedure in 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. Setting up the network structure, argument for training times, and gives default input weights. In the general cases, the default input weight can be random numbers, but it used to be between 0.5 ~ -0.5 or 1 ~ -1.
2. To calculate each node’s output by feed forward method:

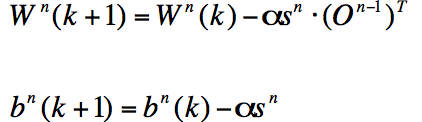


Note that  (input nodes) and  (output nodes).

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



1. Applies to the value O and S came from step 1 and 2, and uses Steepest Descent Algorithm to adjust each nodes’ weights in each layers till convergent. Every time when get a training sample, the system will update and adjust the weights in each nodes.



Repeat the step 2 – 4 until the loop reaches the statement to stop training; A ANN system’s completeness depends on how many training time given, there has two methods to set up for training time: giving a specific numbers to do training, or set a statement to control the system to stop training. However, both methods have its risk to cause the errors happen. In the first method, if the users give not enough times for training, the ANN system can’t get enough samples to adjust the right weights for nodes, in the other words, the ANN system can’t be “full trained” to get a satisfied recognition accuracy. In the second method, if it gives an irrelevant argument, it’s easy to get into the infinity loop and can’t stop from trainings, to overcome this situation, it used to give more than one argument to leave the training loop.

**Conclusion**

Voice Recognition system applies to real life a lot, and it’s also one of hottest issues in the modern society. It’s apparently to see VR applications everywhere, from PDA to smart phone, VR got into human’s life closer; the smart phone is a most commons case in the present days, on 2011, Apple Inc. launched the voice recognition secretary “Siri” to newest smart production, it inspired human’s desire of talking to machine indirectly again. On the next year Google also released the voice recognition search names “Google Search” in the newest smart phone’s OS “Jelly Ben”, it applies to ANN and reduces 25% of recognition faults. Thus it can be seen that VR technique has a good economic benefit for these Technology companies, and its research is valuable to be invested. But the VR system in the present days still have bunch of difficulties need to overcome, for example, the VR system in present days still prefer to a perfect environment. But in the real life still can’t avoid those outer effects exist, languages, pronunciations, noises, even the record machine can cause different speech signals. Though after back propagation algorithm has been released on 1986, it helps ANN to overcome many difficulties, but it still have a problem that ANN needs lot of learning information and calculate time, this still far from the ideal system – a real time recognition. On the contrary, fuzzy logic is another method to imitate human’s neural network, it also has been applied to many fields. It has the advantage with fast learning rate, but lower fault toleration. How to take both advantages from ANN and Fuzzy logic and create a voice recognition system, which has high recognition accuracy, high noises toleration, short reflection time is a valuable topic for future researches.

**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”,

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

[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.