Automatic Speech Recognition for Telephone Voice Dialling in Yorùbá

T.S. Ibiyemi
Dept. Electrical & Electronics Engineering
University of Ilorin, Ilorin, Nigeria

A.G. Akintola
Dept. of Computer Science
University of Ilorin, Ilorin, Nigeria

Abstract

Human Computer Interaction is largely by electromechanical devices. These interaction media are grossly not user's friendly and often risky and life threaten in some applications such as driving and phone dialling. This paper presents our work in telephone auto-dialing in yorùbá language. The speech recognition algorithm used was coded in C language and run on a Pentium duo core 2.6 GHz 2 GB RAM PC with a gsm set and a multimedia headset attached to the PC. The experiments yielded 94% speaker recognition rate, and 82% phone sentence recognition rate.

1. Introduction

Human Computer Interaction, HCI, is largely by electromechanical devices such as keyboard, mouse, joystick, printer, and monitor. These interaction media are grossly not user's friendly and often risky and life threaten in some applications such as driving and phone dialling. A natural and better human-machine interaction to eliminate this fatal risk of driving and phoning is voice dialling of phone. However, telephone voice dialling by anybody who has access to the phone is not good enough. Hence, there is a need to be able to authenticate authorised users of the phone. In order to extend this technology to the grass root, this speaker authentication prior to telephone voice auto-dialling is implemented in Yorùbá language.

Yorùbá language is one of the three dominant local languages spoken in Nigeria by about 22 million people. It is a tonal language, that is, the tone of pronunciation of a yorùbá word determines the meaning of that word. This is unlike non-tonal languages where the spelling of a word suffices to infer its meaning. The problem is further compounded by the fact Yorùbá language is full of homographic words. Homographic words are words with the same spelling but having different meanings depending on their pronunciation tones. Yorùbá language has 25 letter alphabets (a , b, d, e, ẹ, f, g, gb, h, i , j , k, l, m, n, o, ọ, p, r, s, ṣ , t, u, w, y), 7 of them are vowels (*a,e, ẹ, i, o, ọ , u*), the remaining 18 are consonants (b, d, f, g, gb, h, j , k, l, m, n, p, r, s, ṣ , t,w, y) . They are three tone levels, namely, low tone, mid tone, and high tone. The low and high tones are represented by grave accent symbol (`), and acute accent symbol (´) respectively. Mid tone has no accent symbol. Accent symbol , where required, is only allowed on vowel in a yorùbá syllable. The ten yorùbá numerals are (*ofo, enń , eji, ẹta, ẹrin, àrún, èfà, eje, ẹjọ, ẹsan*) equivalent of the ten numerals (0,1,2,3,4,5,6,7,8,9) respectively. Mobile telephone numbers in Nigeria consist of 11 digits drawn from these numerals with the first digit always digit 0 for calls within the country.

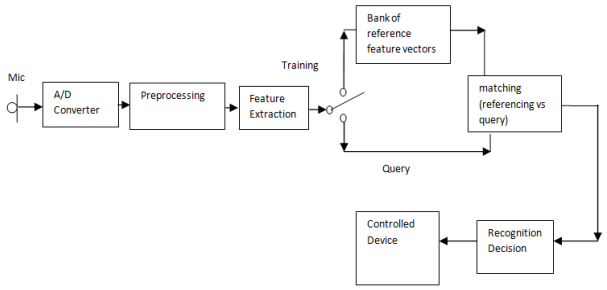
Man-machine interface by speech is a new paradigm shift which is natural and most user's friendly. This new paradigm shift is made possible by automatic speech recognition, ASR, system.

2. Automatic Speech Recognition System

Fig.1 shows the basic structure of an automatic speech recognition system [1, 2]. The air pressure variation caused by speech is transformed to electrical analogue signal by a microphone. The analogue signal output of the microphone is then fed into an analogue to digital converter for conversion into digital samples. These speech samples are pre-processed to put them in a suitable form for easy extraction of the discriminating characteristics inherent in the samples represented as feature vectors. There are two phases to an ASR system, namely, training phase, and query or operational phase. During training phase, the feature vectors that characterized each word or/and speaker are stored as that word's or speaker's reference template in a database.

After the training phase, the system is ready to be deployed for speech recognition. The word to be recognized passes through the microphone, preprocessing and feature extraction units as the training session. However, the extracted feature vectors, this time, are not stored but compared with each word's feature vectors in the reference template

database . The output of the comparison is compared to a threshold for decision making on if the word is recognized or not. The different words used in the training phase constitute the vocabulary of the words that can be recognized.



I ě

3. Hardware Model

The hardware model of this ASR system consists of a PC with an in-built sound card, a microphone, and an rs232 gsm set. A speaker makes an utterance in yorùbá which is captured and traduced from sound-wave to electrical analogue signal by the microphone. The output the microphone is fed into the in-built sound card on the PC for conversion into digital form. During training the digitised speech data are stored for offline processing. The software often used for driving the sound card allows some flexibility in configuration of bitrate, number of audio channels, number of bits per sample. But during the operational phase, speech data is captured online, that is no intermediate storage of speech data is required.

4. Software Model

The software model of the automatic speech recognition consists of series of algorithms for realising the speech recognition as next described [3,4,5,6].

4.1 Pre-processing

The pre-processing implementation consists of the following steps:

(i) Voice Activity Detection, VAD

The VAD, also known as word's endpoint detection, removes the inter-word silence periods. The energy and zero crossing rate, ZCR, of a word is used in conjunction with thresholds to segment the activity area of the word from the silent background. The VAD algorithm is defined as:

Frame Energy:

$$Let\ s_i, i = 1,2,\ldots,N\ be\ a\ frame\ speech\ samples,$$
$$N \Rightarrow no.of\ samples$$
$$E = \frac{1}{N}\sum_{i=1}^{N} s_i^2 \qquad (1)$$

Frame Zero Crossing Rate:

$$z = \frac{1}{N}\sum_{j=2}^{N} \frac{\left|\mathrm{sgn}(s_j) - \mathrm{sgn}(s_{j-1})\right|}{2}$$
$$where: \qquad (2)$$
$$\mathrm{sgn}(s()) = \begin{cases} +1 & , s() \geq 0 \\ -1 & , s() < 0 \end{cases}$$

Energy Upper Threshold:

$$T_u = \frac{1}{L}\sum_{i=1}^{L} E_i \qquad (3)$$

$where$:

$$E_i \Rightarrow i-th\ frame\ energy$$
$$L \Rightarrow no.\ frames\ assumed\ as\ noise, 10\ in\ our\ case$$

Energy Lower Threshold:

$$T_l = 0.25 * T_u \qquad (4)$$

Zero Crossing Rate Threshold:

$$T_z = \frac{1}{L}\sum_{i=1}^{L} z_i \qquad (5)$$

$where$:

$$z_i \Rightarrow i-th\ frame\ zcr$$
$$L \Rightarrow no.\ frames\ assumed\ as\ noise,$$
$$10\ in\ our\ case$$

(ii) Pre-Emphasis Filter

The high pass FIR filter implemented has a transfer function of:

$$H(z) = 1 + az^{-1} \qquad (6a)$$
$$where:\ a = 0.95$$

The time response of this filter is:

$$(n) = s(n) - s(n-1) \qquad (6b)$$

(iii) Frame Blocking

The partitioning of a word samples into short blocks in order to make the signal stationary is based on eqn.7:

$$x_{i,j} = y((K-M)i + j),\ j = 0,1,\cdots,K-1;\ i = 0,1,\cdots,L-1;$$

$$where: \qquad (7)$$

$$N \Rightarrow no.\ of\ samples\ per\ word$$

$$L = \left(\frac{N-M}{K-M}\right) \Rightarrow no.\ of\ frames\ per\ word$$

$$M \Rightarrow no.\ of\ overlapped\ samples\ per\ frames$$

$$K \Rightarrow no.\ of\ samples\ per\ frame$$

(iv) Windowing

Frame blocking using rectangular window as defined in eqn (7) lead to ringing frequency response also known as Gibb's phenomenon as a result of sharp edges. Hence, Hamming window which is a raised cosine window is used to smooth the edges. Hamming window is defined by eqn (8):

$$y(n) = x(n).w(n) \qquad (8)$$

$$where:$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad ,0 \le n \le N-1$$

$$w(n) \Rightarrow Hanmmin\ g\ window$$

$$x(n) \Rightarrow speech\ frame\ samples$$

4.2 Feature Extraction

The characterisation of speech signal data for the purpose of simultaneous recognition of what is said and who said it, that is, speech recognition and speaker verification respectively is most efficiently handled by mel frequency cepstral coefficients, MFCC. This is because MFCC characterisation is very similar to that of the human aural perception. Hence, MFCCs are used as feature vectors in this work for simultaneous speaker authentication and speech recognition [5,6,7,8,9].

The process of obtaining the MFCCs involves transformation of the windowed frame of speech data from time domain to frequency domain and then back to time domain after processing. Firstly, the spectrum magnitude of the windowed speech signal data is obtained on a linear scale frequency by FFT. This output magnitude is converted power magnitude which is convolved with the frequency response magnitude output of filter bank on mel-frequency scale. In order to convert the obtained mel-scaled power spectrum to time domain, an the inverse discrete cosine transform, DCT, is taken. The output of the inverse DCT are the mel frequency cepstral coefficients. The algorithm for obtaining the MFCCs is described in eqn(9) to eqn(13).

Apply DFT to each of the windowed speech signal of eqn(8):

$$Y(n) = \sum_{k=0}^{N-1} w(k)x(k)e^{-j2\pi kn/N} \quad ,n = 0,1,\cdots,N-1 \qquad (9)$$

Get the power spectrum of eqn(9):

$$Y'(n) = \left(\sqrt{\left(Y'_{real}(n)^2\right) + \left(Y'_{imag}(n)^2\right)}\right)^2, \quad = \left(Y'_{real}(n)^2 + Y'_{imag}(n)^2\right) \quad n = 0,1,\cdots,N-1 \qquad (10)$$

Convert power spectrum of eqn(10) in linear frequency to power spectrum in mel-scale frequency:

$$P_{mel}(m) = \sum_{k=0}^{N-1} Y'(k).H(k,m) \quad ,m = 0,1,\cdots,L$$

$$where:$$

$$L \Rightarrow no.\ of\ mel\ filters$$

$$H(k,m) = \begin{cases} 0 & ,f(k) < f_c'(m-1) \\ \dfrac{f(k) - f_c'(m-1)}{f_c'(m) - f_c'(m-1)} & ,f_c'(m-1) \le f(k) < f_c'(m) \\ \dfrac{f(k) - f_c'(m+1)}{f_c'(m) - f_c'(m+1)} & ,f_c'(m) \le f(k) < f_c'(m+1) \\ 0 & ,f(k) \ge f_c'(m+1) \end{cases} \qquad (11a)$$

$$f(k) = f_{\min} + (k-1).\Delta f \quad ,k = 1,2,\cdots,N$$

$$where:$$

$$\Delta f = \frac{f_{\max} - f_{\min}}{N-1} \qquad (11b)$$

$$f_{\min} \Rightarrow Minimum\ speech\ frequency$$

$$f_{\max} \Rightarrow Maximum\ speech\ frequency$$

$$f_c'(m) = f'_{\min} + (m-1).\Delta f' \quad ,m = 1,2,\cdots,L$$

$$where:$$

$$f'(m) = 2595\log_{10}\left(\frac{f(m)}{700} + 1\right), m = 1,2,\cdots,L$$

$$\Delta f' = \frac{f'_{\max} - f'_{\min}}{L-1} \qquad (11c)$$

$$f'_{\min} \Rightarrow Minimum mel\ frequency$$

$$f'_{\max} \Rightarrow Maximum mel\ frequency$$

$$f \Rightarrow speech frequency$$

$$f' \Rightarrow mel speech frequency$$

Obtain the logarithm of the mel scale power spectrum:

$$P'_{mel}(m) = \log_{10}\left(P_{mel}(m)\right) \quad , m = 1,2,\cdots,L \qquad (12)$$

Convert logarithm mel frequency power spectrum to time domain cepstral coefficients (mfcc) using inverse DCT:

$$C_i = \sum_{j=1}^{L} P'_{mel}(j)\cos\left(\pi . i(j-0.5)/M\right)$$
$$i = 1,2,\cdots,M$$
$$where: M \Rightarrow no.\ of\ coefficients, \qquad (13)$$
$$L \Rightarrow no.\ of\ mel-filters$$

4.3 Vector Quantisation, VQ

It is usual in speaker and speech recognition involving multiple utterances of words to generate very large number of feature vectors per word during the training phase. Hence, the total number of feature vectors can easily become unmanageable in term of storage as templates or in matching computation. These two problems can render speech recognition in embedded application unrealisable. Hence, it is imperative to use vector quantisation as data compression method. [3,4,5] . The VQ problem definition is:

$$VQ\ Problem\ Definition:$$
$$Given: T = \{X_1, X_2, \cdots\cdots, X_N\}\ feature\ vectors,$$
$$\&\ no.\ of\ desired\ codevectors\ M$$
$$Find\ C = \{c_1, c_2, \cdots\cdots, c_M\}\ codevectors$$
$$\&\ the\ codevectors'\ s\ partition\ regions$$
$$P = \{s_1, s_2, \cdots, s_M\}\ such\ that\ average$$
$$distortion\ D_{ave}\ is\ \min imised$$

This problem is solved, in our case, using LBG-VQ algorithm of Fig. 2.

_start_LBG_VQalgorithm

{

step0: Codebook Initialisation

-Input N, M, X_i = { $x_{i,1}, x_{i,2},\ldots..,x_{i,k}$} , i=1,2...,N ; k=dimension of feature vector

(N→total no. feature vectors in training set; X → training set feature vectors*

*M→no. of codevectors/vocabulary words *)*

-Calculate 1-codevector codebook:

$$c_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (13a)$$

-Set $\varepsilon = 0.01$

-Set m = 1

-Set $n(h) = 0 \quad , h = 1,2,\cdots,M$

-Calculate distortion, D:

$$D = \frac{1}{N.k}\sum_{i=1}^{N}\sum_{j=1}^{k}\left\| x_{i,j} - c_{1,j}\right\|^2$$

step1: Double Codebook Size by Splitting

$$for\ j = 1\ to\ m\ do$$
$$($$
$$c'_i = (+\varepsilon)c_i$$
$$c'_{m+i} = (-\varepsilon)c_i$$
$$)$$
$$m = 2m$$
$$c_i = c'_i \quad , i = 1,2,\cdots,m \qquad (13b)$$

step2: Distribute feature vectors by clustering

-Set $D' = D$

$$for\ i = 1\ to\ N\ do$$
$$(\ for\ j = 1\ to\ M\ do$$
$$(\ j^* = \arg\min_j \left(\left\| x_i - c_j\right\|^2\right))$$
$$s_{j^*} = x_i\ ;$$
$$n(j^*) = n(j^*) + 1$$
$$) \qquad (13c)$$

step3: Update Centroids/Codevectors

$$c_j = \frac{1}{n_j}\sum_{h=1}^{n_j} s_j(h) \quad , j = 1,2,\cdots,M \qquad (13d)$$

step4: Calculate new Distortion

$$D' = \frac{1}{N.M.k}\sum_{i=1}^{N}\sum_{h=1}^{M}\sum_{j=1}^{k}\left\| x_{i,j} - c_{h,j}\right\|^2 \qquad (13e)$$

$$if\left(\frac{D-D'}{D}>\varepsilon\right) Then\ goto\ step2\ otherwise goto\ step5$$

step5: Repeat until desired number of codewords

$$if\ (m<M)\ Then\ goto\ step1\ otherwise step6$$

step6: Output codebook

$$Output\quad c_j\ ,\ j=1,2,\cdots\cdots,M$$

step7: stop

}_*end_LBG_VQalgorithm.*

ĕıĕ

4.4 Recognition

The recognition phase is implemented by simple Euclidean distance measure and empirically determined threshold as given in eqn(14):

$$d_i=\arg\min\left(\sqrt{R_i^{\,2}-Q^2}\right), i=1,2,\cdots,N$$
$$if\ d_i<T\Rightarrow \text{Re}cognised otherwise$$
$$Not\ \text{Re}cognised \qquad (14)$$

5. Experiment and Result

Speech contains more information than what is said but also includes information on the speaker, accent, gender, and age group. Hence, one single process suffices to authenticate the speaker and to recognise the word. Some experiments were conducted using 20 native Yorùbá speakers to pronounce each of the 13 words in the vocabulary 10 times. These 2,600 words were all used for training. For the recognition phase, the 20 speakers that participated in the training pronounced the telephone sentence (*pè fònú + 11 digit phone number of their choice)* and the word (*gbé fònú*) once in Yorùbá. Each word is sampled at the rate of 8000 samples per second with each sampled quantised into 16 bits. Fig.2. shows speech waveforms of a Nigerian mobile phone number 08034265239 pronounced as a sentence in Yorùbá. The pre-emphasis filter coefficient used is 0.97 at a framed window of 256 samples with 128 samples overlap. The MFCC feature extraction method is used having 20 mel filter bank and 16 dimensional feature vectors.

There are 20 codebooks, with one codebook for each speaker. Each codebook has 13 codebook-lets, each codebooklet represents each word of the vocabulary. A codebooklet contains 10 codevectors representing the 10 utterances per word per speaker. The codebooks, codebooklets, and codevectors are appropriately populated during training. An adaptive threshold is determined for each of the codevector, that is each word utterance, during training. A simple Euclidean distance measure is used in matching a test utterance with the templates in the codebooks. The computed distance is compared with the stored thresholds in determining the speaker and recognition of the spoken sentence for auto-dialling of the gsm set. The system is coded in C language and run on a Pentium duo core 2.6 GHz with 2GB RAM on board. The PC has a gsm set and a multimedia headset attached to the PC. However, the final system will be an embedded front-end interfaced to a gsm set. The experiments yielded 94% speaker recognition rate, and 82% phone sentence recognition rate.

6. Conclusion

A user's friendly human computer interaction based on speech recognition for telephone auto-dialling in Yorùbá was developed. The speech recognition algorithm used was coded in C language and run on a Pentium duo core 2.6 GHz 2 GB RAM PC with a gsm set and a multimedia headset attached to the PC. The experiments yielded 94% speaker recognition rate, and 82% phone sentence recognition rate. Though the system was developed on a PC, the target would be an embedded front-end unit interfaced to a gsm set.

7. Acknowledgement

We acknowledge with great appreciation the generous research and development grant received from Federal Government of Nigeria through the STEP-B project to execute this work.

8. References

[1] Lipeika Antanas, Lipeikiene Joana, Telksnys Laimutis, " Development of Isolated word Speech Recognition", Informatica, vol.13, no.1, 2002, pp. 37-46

[2] E-Hocine Bourouba, et al, "Isolated Words Recognition System Based onHybrid Approach DTW/GHMM", Informatica 30, 2006, pp. 373-384

[3] Allam Musa, "MareText Independent Speaker Identification based on K-Means Algorithm", International Journal on Electrical engineering and Informatics, vol.3, no.1, 2011, pp100-108
[4] Srinivasan A., "Speaker Identification and Verification using Vector Quantisation and Mel Frequency Cepstral coefficients", Research Journal of Applied Sciences, Engineering and technology", vol.4, no.1, 2012, pp. 33-40
[5] Satyahad Singh, and Rajan E.G., "MFCC VQ based Speaker Recognition and its Accuracy Affecting Factors", International Journal of Computer Applications, vol. 21, no.6, 2011, pp.1-6
[6] Kekre H.B., and Vaishali Kulkarni, "Performance Comparison of speaker Recognition using Vector Quantization by LBG and KFCG", International Journal of applications, vol.3, no.10, 2010, pp.32-37
[7] Rashidul Hasan, Mustafa Jamil, Golam Rabbani, Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", Proc. 3rd International Conference on Electrical and Computer engineering, ICECE 2004, 28-30 December, Dhaka Bangladesh, 2004, pp. 565-568
[8] Linde Y., Buzo A., Gray R.M., "An Algorithm for Vector Quantiser Design", IEEE Trans on Communications, vol. COM-28, no. 1, 1980, pp. 84-95
[9] Wael Al-Sawalmeh, Khaled Daqrouq, Omar Daoud, Abdel-Rahman Al-Qawasmi, "Speaker Identification System based Mel Frequency and Wavlet Transform using Neural Network Classifier", European Journal of scientific Research, vol.41, no. 4, 2010, pp. 515-525
[10] Srinivassan A., 2011, "Speech Recognition using Hidden Markov Model", Applied Mathematical Science, vol.5, no. 79, 2011, pp. 3943-3948