

10th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 1-2, 2007

Recognition of Tagalog Alphabets Using The Hidden Markov Model

by

Rolando D. Navarro, Jr.

For additional information, please contact:

Author's name : Rolando D. Navarro, Jr.
Designation :
Affiliation : University of the Philippines, School of Statistics
Address : Diliman, Quezon City, 1101
Tel. no. : (0632) 821-824
E-mail : navarro@ieee.org

Recognition of Tagalog Alphabets Using The Hidden Markov Model

By

Rolando D. Navarro, Jr.

ABSTRACT

Speech recognition is the process of transmuting human speech into text via a computer. An isolated word Tagalog alphabet recognizer is developed using the Hidden Markov Model (HMM). The performance of the model was examined at training and recognition stages. Speech was sampled at 16 kHz and was divided into frames tapered by Hamming window. Each overlapping frames has a window length of 400 samples (25ms) and windowing period is 120 samples (7.5ms). Each window was feature extracted using the 12th order Linear Predictive Coding. Five female and five male speakers were trained, varying the number of Vector Quantization (VQ) codebooks K (ranging from 20 to 40) and HMM states N (ranging from 2 to 10). A 99.0% and 92.0% recognition was achieved for female and male training data when $(K, N) = (32, 6)$ and $(K, N) = (36, 7)$ respectively. On the other hand, using the verification data, the recognition accuracy was 85.5%.

Keywords: Speech Recognition, Hidden Markov Models, EM Algorithm, Tagalog Alphabets.

I. Introduction

Automatic speech recognition (ASR) is the process of transforming speech signals into text form via a computer. It has wide range of applications ranging from a hands-off keyboard, to "voiceprint" IDs and has inspired science fiction writers, directors, and cartoonists that one day people can able to talk with the computer. Unfortunately, the design of the ASR is an extremely difficult task by considering several factors such as co-articulation, unclear spacing between spoken words, and acoustic variability such the difference in vocal-tract size, dialect, speaking style, number of trained vocabulary as well as the communication channel. To make the design more tractable, an Isolated Word Recognition using the Tagalog alphabets considered in this study.

II. Tagalog Phonology

The basic unit of speech that can be spoken distinguishably is called a phoneme (Deller, et., al, 1999). Phonemes can be represented with characters enclosed by //. There are two major classes of phonemes namely vowels and consonants. A relatively open tract produces vowels. On the other hand, a relatively closed vocal tract, resulting in an audible effect on the airflow produces consonants.

The Tagalog phoneme is similar to other languages composed of vowels and consonants. In 1940, the renowned Lope K. Santos adapted the Abakadang Tagalog (Santiago and Tianco, 1991) where he categorically described 21 phonemes as follows:

Vowels: /a/, /e/, /i/, /o/, /u/

Consonants: /b/, /k/, /d/, /g/, /h/, /l/, /m/, /n/, /ŋ/, /p/, /r/, /s/, /t/, /w/, /y/, /ʔ/.

The character enclosed by / / is the alphabet that represent the phoneme with the exception of /ŋ/ which corresponds to the digraph *ng* and /ʔ/ is identified with (') as in *bat'a* and (-) as in *may-ari* does not correspond to a specific alphabet but it is characterized by the closure of the glottis.

Vowel	Phoneme Sequence
A	/a/
E	/e/
I	/i/
O	/o/
U	/u/

Consonant	Phoneme Sequence	Consonant	Phoneme Sequence	Consonant	Phoneme Sequence
B	/ba/	L	/la/	R	/ra/
K	/ka/	M	/ma/	S	/sa/
D	/da/	N	/na/	T	/ta/
G	/ga/	NG	/naŋ/	W	/wa/
H	/ha/	P	/pa/	Y	/ya/

Tagalog alphabets are also classified either a vowel or consonants. Vowel alphabets are pronounced similar to its phoneme. Consonants are a combination of two phonemes: the letter's corresponding phoneme followed by /a/. The alphabet NG is an exception, which is pronounced as /naŋ/.

II. Basic Components Of Speech Recognition

Speech recognition has at least two stages: The feature extraction unit which extracts the significant information that correspond to the identity of sound and the pattern comparison unit that chooses the closest pattern such as the Dynamic Time Warping, Time Delayed Neural Networks, and the Hidden Markov Model (Deller, et. al., 1999).

III. Feature Extraction Unit

Speech is a continuous time signal $x_a(t)$ in both time and amplitude domain. It requires the input signal to be discretized in both time and amplitude domain before it is processed in a digital processor. The sample of speech is obtained at a fixed sampling frequency F_s , which approximates the original signal, that is, $x[n] = x_a(nT_s)$ where $T_s = 1/F_s$ is the sampling period. The sampling frequency is restricted by the Nyquist frequency, which states that if the highest frequency component F_o , and then the signal should be sampled at the rate of at least $2F_o$, that is, $F_s \geq 2F_o$ to avoid signal aliasing (Proakis and Manolakis, 2007)

Speech is pre-emphasized using a pre-emphasis FIR filter to spectrally flatten and minimize the finite precision effects. Although speech signals are non-stationary however, dividing the signal into short duration frames (5ms-25ms) results to an approximately stationary signal. Overlapping frames are used which results to smoother spectral frames (Rabiner and Juang, 1993).

The linear predictive coding (LPC) is an AR time series tool in estimating parameters that characterizes linear time varying systems (Makhoul, 1975). The input signal is an excitation of impulse train for voiced sound and random Gaussian noise for unvoiced sound. The switch rules out the signal that has been excited and scaled by a gain G . The signal enters into an all-pole IIR filter, which determines the characteristic of sound. The transfer function in the z -domain denoted as $H(z)$ of the all pole AR(p) Infinite Impulse Response filter is given as

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}}$$

where $s[n]$ is the windowed signal, $u[n]$ is the excitation signal, G is the zero-frequency gain, $\{a_k\}_{1 \leq k \leq p}$ are the LPC coefficients which can be estimated using the Levinson-Durbin Algorithm (Hayes, 1996).

The parameters used in the feature extraction unit are presented in the table below.

Preprocessing	Quantity
Sampling Rate (F_s)	16 kHz
Time Duration (t)	1 sec
Number of Bits (B)	16 (Pulse Coded Modulation)
Pre-emphasis Filter Transfer Function ($G(z)$)	$G(z) = 1 - az^{-1}$ $a = 0.9375$
Window Length (M)	25 sec (400 samples)
Window Period (L)	7.5 sec (120 samples)
Number of Frames (T)	131
Tapering Window $w(m)$	Hamming Window $w[m] = \left[0.54 - 0.46 \cos\left(\frac{2\pi m}{M}\right) \right] \mathbf{1}_{\{0 \leq m \leq M\}}$
LPC Order (p)	12

Note: The z^{-1} operator in Signal Processing literature (Proakis and Manolakis, 2007) corresponds to the backward B operator in Time Series literature (Brockwell and Davis, 1987).

Spectral properties of a frame can be reduced into a set of codebook indices by means of Vector Quantization. The vector quantizer is trained using the k-means algorithm (Gersho and Gray, 1992). The similarity measure used in this study is based on the Euclidean Distance and the number of VQ clusters under investigation ranges from $K=20$ to $K=40$.

IV. Hidden Markov Model

The Hidden Markov Model (HMM) is a statistical technique introduced by Baum and collaborators (Baum and Petrie, 1966; Baum, et. al., 1970) that is used in the pattern recognition applications. From the nomenclature of the HMM, the identity of the states are unknown only the output observations are known. Besides speech recognition, its applications include handwriting recognition (Aritieres, et. al., 2007; Hu, et. al., 1996), DNA sequence modeling (Krogh, et. al., 1994), and as well financial modeling (Ryden, et. al., 1998). In this study, the observation sequence from the VQ codewords is used to train the HMM and the recognized alphabet is classified by finding the maximum a posteriori probability (MAP) of the observation sequence from the VQ. The HMM is characterized by the following parameters (Wilcox and Bush, 1994):

- A set of N states $\mathbf{S} = \{S_1, S_2, \dots, S_N\}$ and the state at time t given as Q_t and the state sequence is denoted as $\mathbf{O} = \{O_t\}_{1 \leq t \leq T}$.
- A set of K observation symbols $\mathbf{V} = \{V_1, V_2, \dots, V_N\}$, the observation at time t given as O_t , and the observation sequence is denoted as $\mathbf{O} = \{O_t\}_{1 \leq t \leq T}$.
- State Transition Matrix $\mathbf{A} = \{a_{ij}\} \in \mathbb{R}^{N \times N}$ $a_{ij} = P(Q_{t+1} = S_j | Q_t = S_i)$
- Output Distribution Matrix $\mathbf{B} = \{b_{jk}\} \in \mathbb{R}^{N \times K}$ $b_{jk} = P(O_t = V_k | Q_t = S_j)$
- Initial State Distribution $\mathbf{P} = \{p_i\} \in \mathbb{R}^N$ $p_i = P(Q_1 = S_i)$

The HMM is tersely denoted as $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{P})$.

There three problems concerning the HMM that are applied to speech recognition (Rabiner, 1989) which are presented below.

Problem 1: Given $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{P})$ and $\mathbf{O} = \{O_t\}_{1 \leq t \leq T}$, how to compute for $P(\mathbf{O} | \lambda)$?

It can be easily shown by Markovianity and Bayes' rule:

$$P(\mathbf{O} | \lambda) = \sum_{\mathbf{Q}} P(\mathbf{O} | \mathbf{Q}, \lambda) \cdot P(\mathbf{Q} | \lambda) = \sum_{\mathbf{Q}} \left(\prod_{t=1}^{T-1} a_{Q_t, Q_{t+1}} \cdot p_{Q_1} \prod_{t=1}^T b_{Q_t}(O_t) \right)$$

Direct computation of $P(\mathbf{O} | \lambda)$ is computationally expensive. However, a practical method of computing $P(\mathbf{O} | \lambda)$ is by the used forward-backward procedure.

Forward Procedure

Let the forward variable $a_t(i) = P(\{O_k\}_{1 \leq k \leq t}, Q_t = S_i | \lambda)$, which can be computed recursively as follows:

- Initialization Step: $a_1(i) = p_i b_i(O_1) \quad 1 \leq i \leq N$
- Recursive Step: $a_t(j) = \left(\sum_{i=1}^N a_{t-1}(i) a_{ij} \right) \cdot b_j(O_t) \quad 1 \leq t \leq (T-1), 1 \leq j \leq N$
- Termination Step: $P(\mathbf{O} | \lambda) = \sum_{i=1}^N a_T(i)$.

Backward Procedure

Let the backward variable $\beta_t(i) = P(\{O_k\}_{(t+1) \leq k \leq T} | Q_t = S_i, ?)$, which can be computed recursively as follows:

- Initialization Step: $\beta_T(i) = 1 \quad 1 \leq i \leq N$
- Recursive Step: $\beta_t(i) = \sum_{j=1}^N a_{ij} \beta_{t+1}(j) b_j(O_{t+1}) \quad (T-1) \geq t \geq 1, 1 \leq i \leq N.$

Problem 2: Given $? = (\mathbf{A}, \mathbf{B}, P)$ and $\mathbf{O} = \{O_t\}_{1 \leq t \leq T}$, how to choose an “optimal” state sequence?

This problem can be solved by a dynamical programming problem known as the Viterbi Algorithm, and it is used in decoding convolutional codewords in transmitted in noisy communication channels (Wicker, 1995). This algorithm is based on maximizing the probability of the partial sequence up to period t denoted as

$$d_t(i) = \max_{\{Q_k\}_{1 \leq k \leq t}} P(\{Q_k\}_{1 \leq k \leq (t-1)}, Q_t = S_i, \{O_k\}_{1 \leq k \leq t} | ?)$$

$$= \max_{\{Q_k\}_{1 \leq k \leq (t-1)}} P(\{Q_k\}_{1 \leq k \leq (t-1)}, Q_t = S_i | ?) P(\{O_k\}_{1 \leq k \leq t} | ?).$$

moreover, this probability can be computed recursively by virtue of Markovianity as follows:

- Initialization Step $d_1(i) = p_i b_i(O_1) \quad ?_1(i) = 0 \quad 1 \leq i \leq N$
- Recursive Step $d_t(j) = \max_{1 \leq i \leq N} [d_{t-1}(i) a_{ij}] b_j(O_t) \quad 1 \leq j \leq N, 2 \leq t \leq T$
 $?_t(j) = \operatorname{argmax}_{1 \leq i \leq N} [d_{t-1}(i) a_{ij}] \quad 1 \leq j \leq N, 2 \leq t \leq T$
- Termination Step $P^* = \max_{1 \leq i \leq N} [d_T(i)] \quad Q_T^* = \operatorname{argmax}_{1 \leq j \leq N} [d_T(i)]$
- Backtracking $Q_t^* = ?_{t+1} [Q_{t+1}^*] \quad (T-1) \geq t \geq 1.$

Problem 3: How to adjust $? = (\mathbf{A}, \mathbf{B}, P)$ and $\mathbf{O} = \{O_t\}_{1 \leq t \leq T}$ so that $P(\mathbf{O} | ?)$ is maximized?

This is a parameter estimation problem which aims to maximize the likelihood $L(?) = P(\mathbf{O} | ?)$ which can be solved using the EM Algorithm (Dempster, et. al., 1977, MacLachan and Krishnan, 1997) where the missing data are the HMM states. Denote $\hat{?}$ the

updated parameter estimator of interest then the updated estimators known as the Baum-Welch re-estimation formulas are given as

$$\hat{\rho}_i = \frac{a_0(i)\beta_0(i)}{\sum_{j=1}^N a_T(j)} \quad \hat{a}_{ij} = \frac{\sum_{t=1}^T a_{ij}a_{t-1}(i)b_j(O_t)\beta_t(j)}{\sum_{t=1}^T a_{t-1}(i)\beta_{t-1}(i)} \quad \hat{b}_{jk} = \frac{\sum_{t=1}^T a_t(i)\beta_t(i)\mathbf{1}_{\{O_t=V_k\}}}{\sum_{t=1}^T a_t(i)\beta_t(i)}$$

In line with the convergence of the EM Algorithm, Baum-Welch re-estimation formulas guaranteed converge towards the local maxima of $L(?)$.

The HMM problems are applied to recognition of speech as follows (Rabiner and Juang, 1989):

- **HMM Training** – Each of the Tagalog alphabets to be recognized is represented by the HMM. The HMM parameters are estimated using a training observation applying the Baum-Welch re-estimation formulas.
- **HMM Recognition** – The observation sequence \mathbf{O} that is represented by a series of VQ codewords. The observation sequence is used to optimize the state sequence, it is determined by using the Viterbi algorithm for each of the HMM. Then the likelihood of \mathbf{O} is computed from each of the HMM using either the forward or backward procedure. Finally, the HMM that gives the maximum a posteriori probability yields the recognized word.

V. Results

A. Speech Acquisition Methodology

The speaker selected is a fluent speaker of Tagalog with no speaking ailment and in proper disposition. The speaker is asked to pronounce their usual utterance of the alphabet for 1 sec. The speakers employed are limited to adult speakers since their vocal tract is more developed compared to its younger counterparts. In addition, speech was acquired under a controlled environment given as below.

- The environment is confined in a quiet room.
- The same microphone transducer is used for all speech recorded
- The microphone must be at most 3 cm. from the mouth and must be directed normal to the speaker's mouth.

B. Training Phase

Five adult male speakers and five adult female speakers were used for training. Experimentation on HMM and VQ were conducted to find the number of codebooks as well as the number of HMM states that would give the highest overall recognition accuracy using the utterance as test speaker. The number of codebooks k ranges from 20 (which correspond to the number of alphabets) to 40 in multiples of four. On the other hand, the number of HMM states N ranges from 2-10 were tested. The highest accuracy rate recorded for the male and female speakers were given as follows:

	Accuracy	K	N
Male	90.0	32	6
Female	99.0	36	7

It is noted that vowels and semivowels outperform the rest of the consonants. Vowels and semi-vowels have achieved a recognition rate of 90%-100%. On the other hand, some consonants have achieved a recognition rate of below 50%. This discrepancy in performance can be attributed by the following reasons:

- The number of consonants outnumbers the number of vowels (15 versus 5). Taking into account the difference in the nature of vowels and consonants, a consonant utterance is less likely to be recognized than vowels.
- Vowels are spectrally more defined than consonants, which is easily recognized by both human and machine. (Rabiner and Juang, 1993).

C. Verification Phase

The verification phase shall consist of three parts:

- Setup 1: Training Data from same set of utterance (5 males, 5 females)
- Setup 2: Training Data but different set of utterances (5 males, 5 females)
- Setup 3: Verification Data (20 males, 20 females).

The table of recognition accuracies is presented in the table below.

SETUP	Setup 1			Setup 2			Setup 3		
	Alphabet	M	F	Mean	M	F	Mean	M	F
A	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.0	92.5
B	100.0	100.0	100.0	100.0	100.0	100.0	95.0	90.0	92.5
K	100.0	60.0	80.0	80.0	60.0	70.0	75.0	65.0	70.0
D	100.0	100.0	100.0	80.0	100.0	90.0	70.0	85.0	77.5
E	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	95.0
G	100.0	100.0	100.0	100.0	100.0	100.0	85.0	85.0	85.0
H	100.0	100.0	100.0	100.0	100.0	100.0	90.0	90.0	90.0
I	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	97.5
L	100.0	100.0	100.0	100.0	80.0	90.0	80.0	70.0	75.0
M	100.0	100.0	100.0	80.0	80.0	80.0	75.0	75.0	75.0
N	100.0	60.0	80.0	100.0	60.0	80.0	90.0	60.0	75.0
NG	100.0	100.0	100.0	100.0	80.0	90.0	90.0	80.0	85.0
O	100.0	100.0	100.0	100.0	100.0	100.0	95.0	95.0	95.0
P	100.0	60.0	80.0	100.0	60.0	80.0	80.0	50.0	65.0
R	100.0	100.0	100.0	100.0	100.0	100.0	85.0	90.0	87.5
S	100.0	100.0	100.0	100.0	100.0	100.0	90.0	90.0	90.0
T	80.0	60.0	70.0	80.0	80.0	80.0	80.0	70.0	75.0
U	100.0	100.0	100.0	100.0	100.0	100.0	95.0	90.0	92.5
W	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	97.5
Y	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	97.5
Mean	99.0	92.0	95.5	96.0	90.0	93.0	88.8	82.3	85.5

The slight decrease in the recognition accuracy of Setup 2 (95.5%) compared to Setup 1 (93.0%) was anticipated since spectral properties used by the same speaker were common to the test data. On the other hand, the decrease in recognition accuracy of Setup 3 compared to Setup 2 was also anticipated since there could be spectral properties in the test data that were not common to the training data. Training more speakers would enable to capture the spectral properties related to the alphabet. Nevertheless, an overall recognition accuracy of 85.5% is satisfactory for this recognition system.

VI. Conclusion and Recommendations

This study has successfully developed a recognition system that will recognize the Tagalog alphabet using the HMM. The recognition accuracy of 95.5% using the training data and 85.5% using the verification data was quite encouraging. Furthermore, longer-term research must continue beyond the isolated word Tagalog recognition system by extending

this work using the continuous word Tagalog speech recognition and ultimately the development of a commercial Tagalog speech recognition hardware/software.

References

- Aritieres, T. Maruktat, S., and Gallinari, P. (2007). "Online Handwritten Shape Recognition Using Segmental Hidden Markov Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:205-217.
- Baum, L. E. and Petrie, T. (1966). "Statistical Inference for Probabilistic Functions for Probabilistic Functions of Finite State Markov Chains," *Annals of Math. Stat.*, 37:1554-1563.
- Baum, L. E., Petrie, T., Soules, G., and Weinss, N. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Math. Stat.*, 41:164-171.
- Brockwell, P. and Davis, R. (1987). *Time Series: Theory and Methods*, Springer Verlag, New York.
- Deller, J. R. Jr., Hansen, J. H. L., and Proakis, J. G. (1999). *Discrete-Time Processing of Speech Signals*, 2nd ed., Wiley-IEEE Press, Piscataway, NJ.
- Dempster, A. P., Laird, N.M., and Rubin, R. B. (1977). "Maximum Likelihood form Incomplete Data via the EM Algorithm (with discussion)," *J. of Royal Statistical Society B*, 45:51-59.
- Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*, Kluwer, Norwell, MA.
- Hayes, M. (1996). *Statistical Digital Signal Processing and Modeling*, Wiley, New York.
- Hu, J., Brown, M. K., And Turin, W. (1996). "HMM Based On-Line Handwriting Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18:1039-1046.
- Krogh, A., Brown, M. Mian, I.S., Sjölander, and Haussler, D. Hussler (1994). "Hidden Markov Models in Computational Biology: Applications in Protein Modeling," *J. Molecular Biology*, 235:1501-1531.
- Maclachan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Prentice Hall, New York.
- Makhoul, J. (1975). "Linear Prediction: A Tutorial Review," *Proc. IEEE*, 63: 561-580.
- Proakis, J.G. and Manolakis, D.G. (2007). *Introduction to Digital Signal Processing: Principles, Algorithms, and Applications*, 4th ed. New York, Prentice Hall.
- Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, 77:257-286.
- Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, Prentice Hall.
- Ryden, T., Terasvirta, T., and Asbrink, S. (1998). "Stylized Facts of Daily Return Series and the Hidden Markov Model," *J. Applied Econometrics*, 13:217-244.
- Santiago, A.O. and Tianco, N.G. (1991). *Makabagong Bararilang Pilipino*, 3^d ed., Quezon City, Rex Publishing.
- Wicker, S.B. (1995). *Error Control Systems for Digital Communication and Storage*, Prentice Hall, Englewood Cliffs, NJ.
- Wilcox, L. and Bush, M. (1994). *A Handbook of Electrical Engineering*, IEEE Press, New York.