An Automatic Speech Recognition for the Filipino Language using the HTK System

John Lorenzo Bautista, and Yoon-Joong Kim
Department of Computer Engineering, Hanbat National University, Daejeon, South Korea

Abstract - *This paper presents the development of a Filipino speech recognition using the HTK System tools. The system was trained from a subset of the Filipino Speech Corpus developed by the DSP Laboratory of the University of the Philippines-Diliman. The speech corpus was both used in training and testing the system by estimating the parameters for phonetic hmm-based (Hidden-Markov Model) acoustic models. Experiments on different mixture-weights were incorporated in the study. The phoneme-level word-based recognition of a 5-state HMM resulted an average accuracy rate of 80.13 for a single-Gaussian mixture model, 81.13 after implementing a phoneme-alignment, and 87.29 for the increased Gaussian-mixture weight model. The highest accuracy rate of 88.70% was obtained from a 5-state model with 6 Gaussian mixtures.*

Keywords: Speech Recognition; Filipino Language; Acoustic Model; Hidden Markov Model; HTK System;

1 Introduction

Speech is the most effective means of natural communication between humans. This is one of the very first skills humans acquire by learning through active interaction with the environment by mimicking sounds and voices. During the past years, computer scientists and engineers have been eager to incorporate this skill through intelligent machineries to provide a better and effective means of communication between humans and machine.

Developing an Automatic Speech Recognition system (ASR) fills the gap between a more natural means of communication between humans and intelligent systems. ASR systems were designed to recognize speech data and process them into text via natural language processing. These systems have been effectively integrated to the English language, however, only a few study related on the Filipino Language were realized[1]. A cross-training approach was one mean of implementing an appropriate language adaptation for the Filipino language using languages such as English and Indonesian [1][2]. Other studies use a Phonetically-balanced wordlist to train phonetic models for the Filipino Language [3].

In 2002, a Filipino Speech Corpus (FSC) was created by the Digital Signal Processing Laboratory of the University of the Philippines-Diliman [4]. This allowed multiple studies to be conducted relating to the speech recognition of the Filipino Language [5].

In this study, a Hidden-Markov Model (HMM) based ASR system will be used to train and recognize Filipino words using the HTK system[20] by training and testing for a phoneme-level speech recognition system using a subset of the Filipino Speech Corpus

This paper is organized as follows: Section 2 describes the Filipino language as well as the phoneme set used in this paper. Section 3 provides an introduction to the Filipino Speech Corpus (FSC) by the DSP Laboratory of the University of the Philippines. Section 4 shows the Data preparation done for this paper, while Section 5 provides information on the ASR system and the methodology for training data using the HTK System. Section 6 describes the results obtained from the experiments, which is discussed on Section 7 with the conclusions and future works that will be associated with this study.

2 The Filipino Language

The Filipino language is the lingua-franca of the Philippines and is largely used by around 22 million native speakers [7]. The Filipino language is prestige register of the Tagalog Language [8]. Tagalog varies more or less likely from places to places, with distinctive regional dialects noted among the provinces of Bataan, Bantangas, Bulacan, Manila, Tanay, and Tayabas, with the Bulacan dialect considered as the purest, while the Manila dialect is commonly treated as the standard of the Filipino pronunciation[9].

The Filipino language has evolved rapidly within the past few decades. Between 1930s and the mid 1970's a system of syllabication for the alphabet called the abakada was used to represent the native sounds of the language [10].

In a 2013, a new guideline for the language's orthography was published by the Komisyon sa Wikang Filipino (Commission for the Filipino Language)[11]. This guideline emphasized the use of a new set of Filipino alphabet and its phonetic counterpart:

Consonants
b, c, d, f, g, h, j, k, l, m, n, ng, p, q, r, s, t ,v, w, x, y, z

Vowels
a, e, i, o, u

This modern alphabet set was the basis for the phoneme set used in this study. The presentation was patterned on the phonetic code used in the TIMIT database [12], with 35 phonemes consisting of 14 vowels (including diphthongs) and 21 consonants. The phonetic table of the designed phoneme sets is shown as follows:

TABLE I
A PHONETIC CODE FOR THE FILIPINO LANGUAGE USED IN THIS STUDY MODELED AFTER TIMIT

Symbol	Example Word	Syllabication	Transcription
ah	b**A**boy	/ba-boy/	b **ah** b oy
ao	k**O**nti	/kon-ti/	k **ao** n t iy
aw	sig**AW**	/si-gaw/	s iy g **aw**
ax	t**A**nda	/tan-da/	t **ax** n d ah
ay	bant**AY**	/ban-tay/	b ax n t **ay**
b	**B**ihis	/bi-his/	**b** iy h ih s
ch	**CH**okoleyt	/cho-ko-leyt/	**ch** oh k oh l ey t
d	**D**ilaw	/di-law/	**d** iy l aw
eh	**E**spiritu	/es-pi-ri-tu/	**eh** s p iy r iy t uh
ey	r**EY**na	/rey-na/	r **ey** n ah
f	**F**iesta	/fi-yes-ta/	**f** iy y eh s tah
g	**G**ulay	/gu-lay/	**g** uh l ay
h	**H**ayop	/ha-yop/	**h** ah y ao p
ih	**I**ntay	/in-tay/	**ih** n t ay
iw	sis**IW**	/si-siw/	s iy s **iw**
iy	**I**bon	/i-bon/	**iy** b ao n
j	**J**eepney	/jip-ni/	**j** ih p n iy
k	**K**amay	/ka-may/	**k** ah m ay
l	**L**arawan	/la-ra-wan/	**l** ah r ah w ax n
m	**M**ata	/ma-ta/	**m** ah t ah
n	**N**ais	/na-is/	**n** ah ih s
ng	**NG**ayon	/nga-yon/	**ng** ah y ao n
oh	**O**ras	/o-ras/	**oh** r ax s
ow	epis**OW**d	/e-pi-sowd/	eh p iy s **ow** d
oy	tul**OY**	/tu-loy/	t uh l **oy**
p	**P**ula	/pu-la/	**p** uh l ah
r	**R**egalo	/re-ga-lo/	**r** eh g ah l oh
s	**S**ipa	/si-pa/	**s** iy p ah
t	**T**ao	/ta-o/	**t** ah oh
ts	**TS**inelas	/tsi-ne-las/	**ts** iy n eh l ax s
uh	**U**bo	/u-bo/	**uh** b oh
v	**V**igan	/vi-gan/	**v** iy g ax n
w	**W**ika	/wi-ka/	**w** iy k ah
y	**Y**aman	/ya-man/	**y** ah m ax n
z	**Z**igzag	/zig-zag/	**z** ih g **z** ax g

3 The Filipino Speech Corpus

The Filipino Speech Corpus (FSC) was created by the Digital Signal Processing Laboratory (DSP Lab) of the UP Electrical and Electronics Engineering Institute with the help of the UP Department of Linguistics in 2002 [2]. The corpus contains more than 100 hours of read Filipino text prompts and spontaneous Filipino speech. The corpus was recorded by 50 female and 50 male speakers from different parts of the Philippines using high-fidelity audio recording equipment. A variation to the FSC was created in 2009 that included syllabic utterances (bi-phones). This recent version is used in this study.

4 Data Preparation

In this section, we will discuss the data preparations done for the resources used in the speech recognition system including the development of a lexicon and pre-processing of the speech data.

4.1 Lexicon Development

An 18,000 unique word lexicon was developed by the Intelligent Information and Signal Processing Laboratory (IISPL) of Hanbat National University. This lexicon is labeled as “Hanbat Filipino Lexicon” was based on words gathered from several wordlists on the internet including entries from a medium-sized tri-lingual dictionary [14]. The lexicon consists of the word entry id, its corresponding written format, and the associated pronunciation as follows :

SIKAPING [SIKAPING] s iy k ah p ih ng
SIKAT-adj [SIKAT] s iy k ax t
SIKAT-nn [SIKAT] s ih k ax t
…

Homographs with different pronunciations were associated with an additional token for distinguishing the differences between the words while homophones with the same pronunciation were treated as one.

This lexicon was used to convert the text transcriptions from the FSC to its phonetic transcription patterned on the TIMIT phoneme set, as discussed in section 2.

A total of 1913 Out-of-Vocabulary (OOV) words were found from the FSC after processing with the Hanbat Lexicon (including 431 bi-phone utterances). These OOV words were manually transcribed and were included into to a separate lexicon.

4.2 Preprocessing the Speech Data

The FSC contain speech data recorded at 44.1 kHz for each recording session. These data were recorded continuously and are saved in one file (waveform format). This became inconvenient for the researchers to process the data especially with the unintelligible utterances and noises that were included in the data.

The data were segmented using the program Transcriber [15] based on the specified transcription files from FSC. The data were then pre-processed by removing the unintelligible utterances and noises, as well as reducing the silence parts of each sentence start and ending, programmatically. Mistakes in the transcriptions as well as incorrect timing were fixed in this phase.

Only a subset of the FSC were used in the training and testing of the system (prior to the 2009 variation) which includes 20 male and 20 female speakers from the total population of 50 male and 50 female speakers.

5 Building the Acoustic Model using HTK

The HTK system is based on Hidden Markov Models (HMM) [20]. HMMs are probabilistic models used for modeling stochastic sequence with underlying finite state structure. This model is widely used to model speech utterances in contemporary speech recognition development and is used by multiple ASR adaptations in languages such as Vietnamese, Indian, Polish, etc. [16][17][18].

In a phoneme based ASR, the HMM would represent the set of phonemes of the language, with each phoneme associated with a single HMM model [19]. A phoneme model (w) - denoted by HMM parameters (lambda) - is presented with a sequence of observations (sigma) to recognize a phoneme with the highest likelihood given:

$$w^* = \arg\max(w|W)P(\sigma|\lambda_w) \quad (1)$$

where:

w = phoneme
W = total phoneme set
σ = observation values
λ_w = HMM model for phoneme w

In which the model λ_w visits different states x_t to generate an output o_t given time *t*. The doubly stochastic processes allows the state transitions and output generation depending on the actual state. The model is defined with a set of N reachable states $S = \{s_1, s_2, \ldots, s_N\}$, and a set of M possible output symbols $V = \{v_1, v_2, \ldots, v_M\}$. A set of three possible measures, *A*, *B*, and π is required to specify an HMM model λ as $\{ A, B, \pi \}$ where :

$A = \{a_{ij}\}$ as the state transition probability
$B = \{bj(k)\}$ as the observation probability
$\pi = \{\pi_i\}$ as the initial state distribution

5.1 Acoustic Processing

The speech data were encoded into Mel-frequency cepstral coefficient (MFCC) vectors. This vector values represents the speech signals' power spectrum (frequency content of the signal over time). Each vector values is the peak of the power spectrum called a formant frequency.

To compute for the format frequency, a series of signal processing are implemented to the speech signal. A Discrete Fourier Transform (DFT) is used to compute the frequency information in the time-domain, and is processed using a Fast-Fourier Transform (FFT) to contain real point values for the magnitude and phase information from the time domain.

Features from these values were extracted using Mel-cepstral filter banks, where in the central frequencies and bandwidth were scaled, and the FFT values were filtered based on the each Mel filter bank. These coefficients correlate to the formant frequencies of the speech signal. All these processes were obtain using the HTK tool HCopy which converts the speech file to its MFCC values.

The MFCC values computed includes the First Derivative and Second Derivative of the 13 MFCC values (12 MFCC + 1 zeroth Coefficient), totaling to 39 Vector Values.

The coding parameters include the use of 13 MFCC Features (12 MFC Coefficients + 1 Zeroth Coefficient/Energy), MFCC Derivatives, and MFCC Acceleration values. A Pre-emphasis coefficient of 0.97 is for the pre-processing, while a Hamming windowing is used in the experiments with a window size of 25ms per window. 26 Filterbank channels are used with 22 Cepstral Lifters.

5.2 Prototyping the HMM Models

The speech data were trained in a phoneme-based HMM model in 4, 5, 6, and 7 states for each phoneme, with the first and last states consisting of non-emitting start and end states. An HMM topology for a 6-state

An initial model (prototypes) for each n-state models were initialized for the training. The HTK tool *HCompV* was used to compute for the global means and variances for all Gaussians used in the HMM. These values are used for the prototype model that used in the embedded training of the model. The embedded re-estimation uses a Baum-Welch Re-estimation algorithm to compute for the new hmm parameters using the HTK tool *HRest*. Each n-state models were estimated for 5 iterations before being aligning the phonemes in the data.

A total of 37 models are used consisting of the 35 phonetic models as well as 2 silence models (a silence model “SIL” for long silences, and a short pause model “SP” for shorted silences or breaks).

A phoneme alignment is done to fit the right phoneme utterances for word that have more than one pronunciation. Alignment provides a new set of speech transcription for better training of the phoneme models. This phoneme alignment is done using the HTK tool *HVite*. Once the phonemes were aligned, each n-state models were re-estimated for an additional 15 iterations.

5.3 Refining the Phonetic Models

A single phonetic model (monophone model) has some problems in terms of co-articulations of two consecutive phonemes. This co-articulation problem could be addressed by increasing the phoneme models from monophones to triphone models. However, the data used in this study does not contain enough training data to provide accurate triphone-based acoustic models. To improve the monophone models, a fine-tuning of the models are used to refine the system by increasing the Gaussian mixtures. Initially, for the first 20 iterations, only one Gaussian mixture is used.

To refine the monophone models, an additional 12 iterations for each n-state models are performed while increasing the Gaussian mixture weights by increments of 2 for every two consecutive iterations.

6 Testing and Results

The performance of the ASR’s trained acoustic models were tested with a 500 word isolated speech data. These data were tested for each iterations of HMM parameters for iterations 6 to 32. Results of the testing are shown in Table 2, with the highest accuracy rate of **81.52%** for the 5-state model in a single Gaussian mixture.

TABLE II
Maximum Recognition Rate for each n-State model for the 20 re-estimation of the HMM models

States	Recognition Accuracy Rate (%)
4	77.22
5	81.52
6	77.07
7	62.50
Average	74.58

The results imply that the 5-state model provides the acoustic model representation for the phoneme-sets based on the training data used. The average recognition rate of the n-state models of this study is 74.58%.

TABLE III
Accuracy Rate for Tests before Phoneme Alignment
(Iterations 6 to 10)

Iterations	Number of States			
	4	5	6	7
6	74.74	80.32	75.45	63.31
7	75.78	81.12	76.83	64.27
8	76.78	81.85	77.68	64.81
9	77.63	82.61	78.46	65.18
10	75.15	79.65	75.8	62.43
Average	76.15	80.13	76.03	61.91

Table 3 shows the results of the testing data based on the HMM parameters estimated for iterations 6 to 10, without phoneme alignment. Table 4 on the other hand shows the results based on the HMM parameters estimated for iterations 11 to 20 after phoneme alignment. Results for the experimental test shows that the highest recognition rate was achieved at the 9th iteration without phoneme alignment.

TABLE IV
Accuracy Rate for Tests after Phoneme Alignment
(Iterations 11 to 20)

Iterations	Number of States			
	4	5	6	7
11	75.55	79.70	75.89	62.00
12	76.03	80.03	75.89	61.91
13	76.37	80.26	76.11	61.85
14	76.63	80.51	76.22	61.89
15	76.81	80.70	76.43	62.07
16	76.96	81.05	76.58	62.15
17	77.06	81.15	76.71	62.33
18	77.14	81.27	76.89	62.41
19	77.24	81.43	77.05	62.51
20	77.22	81.52	77.07	62.50
Average	77.02	81.13	76.74	62.30

The increase in recognition rate based on the number of re-estimations for each n-state models for iterations

6 to 20 using a single Gaussian mixture weight are shown in Figure 2.

TABLE V
ACCURACY RATES FOR EACH N-STATE
MODELS AFTER INCREASING THE GAUSSIAN MIXTURE WEIGHTS

Mixtures	Re-estimations	Number of States			
		4	5	6	7
2	21	78.78	82.74	78.51	64.06
	22	81.79	84.80	80.60	68.03
4	23	84.78	85.80	81.37	70.20
	24	87.69	87.29	82.95	73.32
6	25	88.65	88.27	83.92	74.89
	26	88.62	88.70	83.82	75.23
8	27	88.48	88.51	84.09	75.38
	28	87.72	88.21	84.59	75.00
10	29	87.81	88.26	84.35	74.72
	30	88.28	88.20	84.23	74.43
12	31	88.86	88.44	84.02	74.51
	32	88.86	88.26	84.96	74.42
Average		86.69	87.29	83.12	72.85

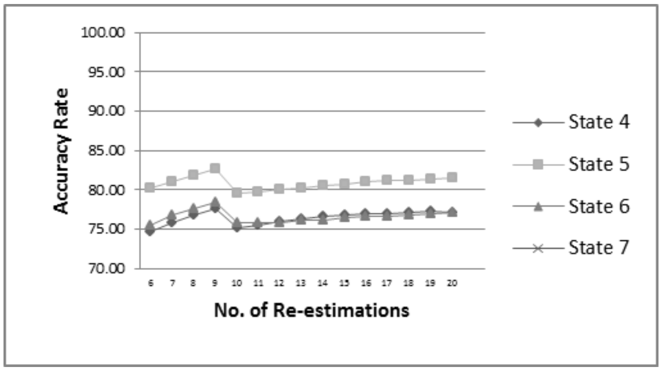

Figure 2. Accuracy Rate for each re-estimation with a single Gaussian mixture

After Increasing the Gaussian Mixture Weights for the n-state models, the testing data were again tested against the HMM parameters for the iterations 21 to 32. Accuracy Rates for the mixture systems are shown in Table 5.

Table 5 shows the results based on the increased Gaussian mixtures models. Results for the experimental test shows that the highest recognition rate was achieved at 6 Gaussian mixtures for a 5-state model with an accuracy rate of **88.70%.**

A graphical representation of the recognition rates for the mixture systems (iterations 21 to 32) are shown in Figure 3.

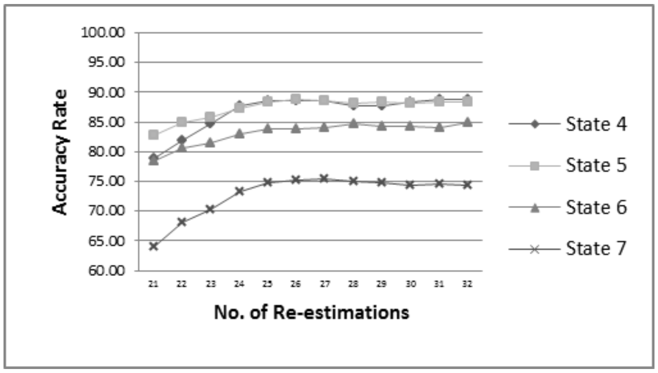

Figure 3. Accuracy Rate for each re-estimation with increased Gaussian Mixture Weights

7 Conclusion and Future Works

In this study, we have presented a speech recognition based on HTK for the Filipino Language using the Filipino Speech Corpus (FSC) by the University of the Philippines-Diliman. A subset of the FSC was taken to be trained using the HTK System tools with recordings by 20 male and 20 female speakers. The data were pre-processed and cleaned by removing non-intelligible sounds and noises, as well as reducing the silences in the wave file. The transcriptions were converted to its phonetic transcription using an 18,000 word lexicon developed by IISPL in Hanbat National University. The data were trained at 4, 5, 6, and 7-state HMM models. The experiments included a single-Gaussian mixture model training, incorporating phoneme-alignment, and increased Gaussian mixture weights.

Results from the experiment showed that a 5-state HMM model is the best n-state model based on the experiments, with an average accuracy rate of **80.13** for a single-Gaussian mixture model, **81.13** after implementing a phoneme-alignment, and **87.29** for the increased Gaussian-mixture weight model. While the highest accuracy rate of **88.70%** for a 5-state model with 6 Gaussian mixtures.

4 and 5 state models are recommended for applications for speech recognition based on the experiments performed in this study, with the lesser n-state model (4-state) to be more robust and faster compared to higher n-state models. A significant drop to the accuracy rates were also noticed for the 6 and 7-state models as compared to the 5-state model. A significant increase in the accuracy rate after increasing the Gaussian mixtures noted. However, increasing the mixture weights would also affect the robustness of the recognition system because of the increased number of computations performed compared to a single Gaussian mixture model.

For future research, the training data will be increased with the full FSC data which will be tested with a tri-phone based HMM model as well as incorporating a language model to improve the recognition capability of the system with sentences, and increasing the developed Filipino lexicon into a larger word entry count. As of the moment, a Filipino language model is underworks and is expected to be included on the future research experiments.

8 Acknowledgment

The researchers would like to thank the Digital Signal Processing Laboratory (DSP Lab) of University of the Philippines-Diliman for providing the Filipino Speech Corpus (FSC).

9 References

[1] Sagum, R., Ensomo, R., Tan, E., Guevara, R.C., “Phoneme Alignment of Filipino Speech Corpus”, *Conference on Convergent Technologies for Asia-Pacific Region, TENCON-2003*, October 2003

[2] Sakti S., Isotani, R., Kawai H., and Nakamura, S.,” The Use of Indonesian Speech Corpora for Developing a Filipino Continuous Speech Recognition System*”,* 2010

A. Fajardo, Y. Kim, “Development of Fillipino Phonetically-balanced Words and Test using Hidden Markov Model”, *Proc. International Conference on Artificial Intelligence*, pp. 22-25, United States of America, July 2013

[3] Guevara R., Co, M., Espina, E., Garcia, I., Tan, E., Ensomo R., Sagum R., “Development of a Filipino Speech Corpus”, *3rd National ECE Conference, Philippines,* November 2002

[4] Guevara R., et. al., “Filipino Databases and their Applications to Educational Institutions”, *1st Philippine Conference Workshop on Mother Tongue-Based Multilingual Education,* February 2010

[5] Rara, K., Cristobal, E.R., de Leon, F., Zafra, G., Clarin, C., Guevara R., “Towards the Standardization of the Filipino Language: Focus on the Vowels of English Loan Words”, *2009 International Symposium on Multimedia and Communication Technology (ISMAC 2009),*Bangkok, Thailand, January 2009

[6] http://wika.pbworks.com/w/page/8021671/Kasaysayan, “Ebolusyong ng Alpabetong Filipino”, Retrieved 2014

[7] Nolasco, R., “Ang Filipino at Tagalog, Hindi Ganoong Kasimple” (*Filipino and Tagalog, it’s not that simple),* Komisyon sa Wikang Filipino, 2007

[8] Schachter, P., Otanes, Fe., “Tagalog Reference Grammar”, *University of California Press,* (1972)

[9] http://wika.pbworks.com/w/page/8021671/Kasaysayan, “*Ebolusyong ng Alpabetong Filipino*”, Retrieved 2012

[10] Almario, V., “Bagong Gabay sa Ortograpiya ng Wikang Filipino” (*New Guidelines for Orthography in the Filipino Language)*, Komisyon sa Wikang Filipino, 2013

[11] Garofolo, J., et. al., “TIMIT Acoustic-Phonetic Continuous Speech Corpus”, *Linguistic Data Consortium*, Philadelphia, 1993

[12] Lazaro R., Policarpio L., Guevara R., “Incorporating Duration and Intonation Models in Filipino Speech Synthesis,” *2009 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009)*, October 2009

[13] S. Calderon, “Diccionario Ingles-Español-Tagalog”, Manila, Philippines, 2012

[14] Barras C., et. al, “Transcriber: a Free Tool for Segmenting, Labeling, and Transcribing Speech”, *First International Conference on Language Resources and Evaluation (LREC)*, May 1998

[15] Nguyen, H., et. al., “Automatic Speech Recognition for Vietnamese Using HTK System”, *International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, Hanoi, November 2010

[16] Al-Qatab, B., et. al, “Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK)”, *International Symposium in Information Technology (ITSim),* Kuala Lumpur, June 2010

[17] Zi'olko, B., et. al., “Application of HTK to the Polish Language”, *International Conference on Audio, Language and Image Processing ICALIP 2008,* Shanghai, July 2008

[18] Rabiner, L. R. and Juang, B. H., “*Fundamentals of Speech Recognition”*, Englewood Cliffs, NJ, Prentice Hall, 1993

[19] S. Young, “Hidden Markov Model Toolkit: Design and Philosophy”, *CUED/F-INENG/TR.152*, Cambridge University Engineering Department, September 2012