G2P and ASR techniques for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon

Angelina Aquino, Joshua Lijandro Tsang, Crisron Rudolf Lucas, and Franz de Leon
Digital Signal Processing Laboratory, Electrical and Electronics Engineering Institute
University of the Philippines, Diliman, Quezon City, Philippines
{angelina.aquino, joshua.tsang, crisron.lucas, franz.de.leon}@eee.upd.edu.ph

***Abstract*—Philippine linguists are tasked with documenting over 170 indigenous languages. A key part of this documentation is the phonetic transcription of recorded speech, which is typically done by hand, and is often expensive and time-consuming. Automated phonetic transcription systems provide a faster and cheaper alternative to manual transcription, but no such system has yet been developed for most Philippine languages. In this paper, we present an implementation of three APT methods—grapheme-to-phoneme conversion, automatic speech recognition, and adaptive alignment—for transcription of small speech corpora in Tagalog, Cebuano, and Hiligaynon. We show that the G2P, adaptive, and select ASR models perform at par with human transcribers while greatly reducing total time and costs. These systems serve as a competent baseline for future developments in APT for Philippine languages, and are expected to facilitate further research and advancements in Philippine linguistics and speech technology.**

***Index Terms*—phonetic transcription, grapheme-to-phoneme conversion, automatic speech recognition, low-resource languages**

I. Introduction

Phonetic transcription is the process by which spoken language is represented based on its constituent speech sounds, called phones, which are each mapped to unique symbols in text. These transcriptions are necessary in documenting and characterizing different languages, as well as developing applications for speech recognition and synthesis. As the Philippines is home to more than 170 indigenous languages [1], phonetic transcription in these languages is vital to continuing developments in linguistics and multilingual education. However, the manual transcription process entails considerable time and expense, with trained linguists requiring 3-4 hours of labor to transcribe 1 hour of audio data, and professional fees being quoted at Php1000 per work hour.

A faster and cheaper alternative to this process is the use of automated phonetic transcription (APT) systems, which enable computers to analyze and annotate speech data at much higher speeds than human linguists. APT systems typically recognize phones from acoustic speech using either generative models (e.g. hidden Markov models), which learn the most likely sequence of phones from a series of acoustic and linguistic observations, or discriminative classifiers (i.e. neural networks), which pass features of speech data through a series of weighted activation functions and produce a classification at the output [2]. When trained and tested on the TIMIT corpus, a widely-used American English speech database containing five hours of phonetically-annotated speech files, generative models have achieved a baseline 66% phone recognition accuracy as early as 1989 [3], while recent neural network-based systems have reported a record accuracy of 85% on the same corpus [4], [5].

Within the Philippine setting, however, APT systems have only been developed for the Tagalog language, with a peak accuracy of only 67% [6], [7]. These systems do not perform at par with human linguists, whose inter-transcription agreement rates average between 75% and 85% accuracy [8]. Hence, there is a need to improve upon APT performance for Tagalog, and to create systems capable of transcribing speech in other Philippine languages as well.

Unfortunately, unlike in English and other well-documented languages, phonetically-annotated audio data on which to train such APT systems are not readily available for Philippine languages. One means to overcome this is with the use of adaptive APT, wherein word-level annotations of the speech files are first converted to phonetic transcriptions, which are then aligned and corrected using the corresponding audio data. Adaptive systems have been proven effective for the Basque language, at 80% accuracy [9], and Austrian German, with accuracies of up to 98% in read speech and 82% in conversational speech [10].

In this paper, we present an adaptive APT system for transcription of small speech corpora in Tagalog, Cebuano, and Hiligaynon, three of the four most spoken languages in the Philippines. Our system consists of three subsystems:

1) a grapheme-to-phoneme (G2P) converter, which parses the word-level text transcription of a given speech utterance and returns the most likely phonetic equivalent of each word;
2) an automatic speech recognition (ASR) tool, which processes the audio recording of a speech utterance, and returns the sequence of recognized phones; and
3) an adaptive substitution phase, which aligns the two prior transcriptions using an optimal matching algorithm, modifies the G2P transcription if a permissible allophone is detected using the ASR system, and returns the adapted phonetic transcription.

We evaluate these systems based on the phone error rates (PER) of their resulting transcriptions, and compare their performance to known inter-transcriber agreement rates. We also report the total running times for training and testing each model across varying durations of input speech data, and discuss the viability

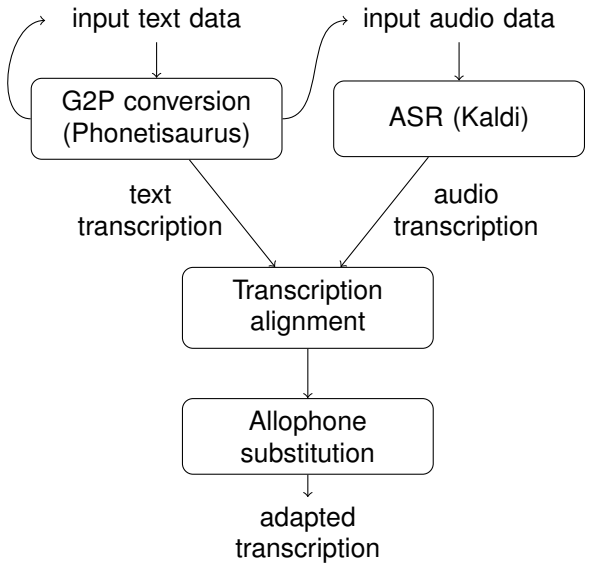

Fig. 1. Block diagram of the APT system

of the three methods as satisfactory alternatives to manual transcription.

II. Speech Corpora for Philippine Languages

Speech corpora for Tagalog, Cebuano, and Hiligaynon were obtained from the ISIP Project 6 Database of the UP Digital Signal Processing Laboratory. An inventory of the phones in the three target languages was then compiled as shown in Table I, including native speech sounds listed in reference grammars for each language [11]–[13], as well as English phones to account for vocabulary and loan words commonly encountered in Philippine speech.

The selected speech corpora consist of recorded audio files and their accompanying word-level text transcriptions. Phonetic transcriptions of these corpora were created using a feedback process: a small number of text files were first hand-transcribed phonetically and used to train a G2P model, which was then used to generate initial phonetic transcriptions of the remaining text and audio files. These transcriptions were manually corrected and verified by representatives of the UP Department of Linguistics to serve as ground truth data for the training and evaluation of the APT system. A summary of the transcribed data, including speaker IDs and corresponding durations, is provided in Table II. The recordings were divided into two

TABLE I
IPA Phone Inventory for Target Languages

Vowels			Consonants					
a	ɛ	i	p	b	t	θ	d	ð
o	u	aɪ	k	g	ʔ	m	n	ɲ
aʊ	ɛɪ	iʊ	ŋ	r	ɹ	f	v	s
oɪ	oʊ	uɪ	z	ʃ	h	ʧ	ʤ	w
ə	æ	ɔ	j	l	ʎ			

TABLE II
Summary of Transcribed Speaker Data

Cebuano		Hiligaynon		Tagalog	
ID	**Length**	**ID**	**Length**	**ID**	**Length**
1691	34m 52s	0738	26m 18s	0156	17m 2s
2295	38m 10s	0739	25m 12s	0812	18m 35s
2998	25m 46s	0740	25m 55s	2359	17m 42s
4703	8m 15s	0741	14m 40s	4281	19m 20s
8973	27m 46s	0742	14m 45s	5093	17m 47s
		0744	29m 5s	5546	20m 40s
				6093	16m 37s
				7859	20m 54s
				8125	17m 58s
				8453	16m 43s
Total	2h 15m	**Total**	2h 16m	**Total**	3h 5m

allocations: Set 1 was limited to 1 hour each per language, while Set 2 contained the entire collection of transcribed data. Aproximately 80% of the speech utterances per set were allocated for training of models, while the remaining 20% were used to test the accuracy of each system.

III. Grapheme-to-phoneme Conversion

The word-level transcriptions of utterances in the speech corpora were stored in formatted log files, while the verified manual transcriptions were encoded in spreadsheets and converted to comma-separated value (CSV) files. Scripts were created in Python to parse these files simultaneously and create dictionaries for each training instance, containing a list mapping all words in the training subset to their corresponding pronunciation variants. An excerpt from the generated training dictionaries is shown in Fig. 2, with lines 7 and 8 demonstrating the possible occurrence of more than one pronunciation for a single word spelling.

The Phonetisaurus toolkit was trained using these dictionaries to generate weighted finite-state transducer (WFST) pronunciation models, which are capable of producing the most likely phonetic transcription of both known and unknown words. Phonetic transcriptions for the test subsets were then generated by applying the pronunciation model to each word in the test log files, and concatenating the pronunciations to produce a single transcription per utterance.

```
1  ang          q a ng
2  angel        q ey n jh e l
3  angeles      q a ng e l e s
4  anggulo      q a ng g o l o
5  anghel       a ng h e l
6  ani          q a n i
7  anibersaryo  a n i b er s a r i y o
8  anibersaryo  q a n i b e er s a er y o
```

Fig. 2. Excerpt from an encoded pronundictionary for G2P training

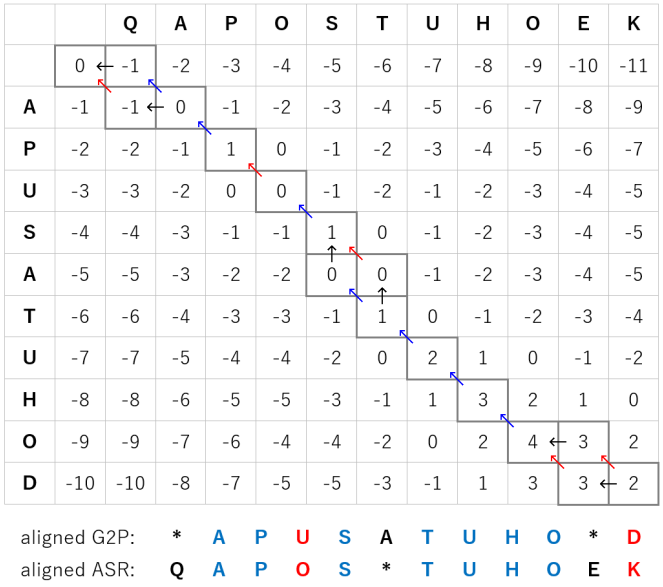

Fig. 3. Sequence alignment with Needleman-Wunsch for $s = \pm 1, W_1 = 1$. Diagonal arrows indicate matches or substitutions in the aligned sequences, while horizontal and vertical arrows indicate insertions and deletions.

IV. Automatic Speech Recognition

The ASR system was created using the Kaldi toolkit [14]. Data for training the acoustic and language models were first prepared. Data prepared for the acoustic model included the audio files and their corresponding transcriptions, speaker information, and a lexicon of all phones to be recognized. The SRI Language Modeling Toolkit (SRILM) was used to build the language model. As this system was designed to distinguish individual phones instead of words, the lexicon was encoded to contain only the inventory of phones as listed in Table I. Likewise, the utterances within the corpus were transcribed in phones instead of words.

A Gaussian mixture model–hidden Markov model (GMM-HMM) framework was used for acoustic modeling while a bigram model was chosen for language modeling after its performance was compared with a unigram model. Using a bigram model, the probability of detecting a certain phone was dependent on the previously detected phone, as opposed to a unigram model with phone probabilities independent of previously detected phones. The acoustic modeling process was better optimized by cycling through training and alignment phases.

The ASR was then trained using monophone and triphone models, as well as Mel-frequency cepstral coefficients with delta feature computation (MFCC+Δ+$\Delta\cdot\Delta$), and linear discriminant analysis with maximum-likelihood linear transform estimation (LDA+MLLT) models. The resulting transcriptions with the highest accuracy were then selected for use in the adaptive phase.

V. Alignment and Adaptive Substitution

A. Transcription alignment

After the speech corpora were processed by the trained G2P and ASR phases, two transcriptions were generated for each utterance. These pairs of transcriptions were prepared for the adaptive substitution phase by aligning each of them using the Needleman-Wunsch algorithm [15], which determines the optimal global alignment of two sequences $A = a_1, a_2, ..., a_n$ and $B = b_1, b_2, ..., b_m$ by incrementally scoring the similarity of their subsequences. Using this algorithm, a score was given to each pair of phones to denote the similarity of a subsequence ending in that pair, with a more positive score indicating higher similarity. This score was calculated by:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ H_{i-1,j} - W_1, \\ H_{i,j-1} - W_1 \end{cases} \qquad 0 \le i \le n, 0 \le j \le m$$

where $H_{i,j}$ is the score for a pair of elements a_i and b_j, $s(a_i, b_j)$ is the substitution score for the two elements (positive if a_i and b_j are identical, negative otherwise), and W_1 is the penalty for a deletion or insertion (otherwise known as the gap penalty). For this project, an arbitrary substitution score of ± 1 and gap penalty of 1 were applied, and the resulting scores were stored in a two-dimensional matrix. The optimal alignment of each pair of transcriptions was found by tracing back the path of adjacent elements with maximum scores, starting and ending at the highest- and lowest-indexed elements of the scoring matrix, respectively. This path returned an alignment which minimizes the edit distance (i.e. the number of deletions, insertions, and substitutions) between the two transcriptions. An example of the scoring matrix and resulting sequence alignment for two transcriptions of a Tagalog utterance is shown in Fig. 3.

B. Allophonic substitution

Once the G2P and ASR transcriptions were aligned, the phonemic G2P transcription was then adapted to account for possible allophonic variations detected by the ASR. A list of observed phonological rules in the target languages was compiled as shown in Table III, specifying possible substitutions, insertions, or deletions, and the environments in which they can occur. Aligned pairs of phones were then checked. If a mismatch was detected, and a substitution between the G2P and

TABLE III
Phonological Rules in Target Languages

Phonological Process	Environment
ɛ ⇄ i	all
o ⇄ u	all
oɪ ⇄ uɪ	all
ɾ ⇄ ɹ	all
∅ ⇄ ʔ	#_V, V_V, V_#
ʧ ⇄ ti(j)	all
ʤ ⇄ di(j)	all
ʃ ⇄ si(j)	all
ɲ ⇄ ni(j)	all

TABLE IV
SYSTEM PERFORMANCE FOR SINGLE-LANGUAGE AND MULTILINGUAL DATASETS

	Cebuano		Hiligaynon		Tagalog		Multilingual	
	1h	2h	1h	2h	1h	3h	3h	7h
G2P accuracy	93.80%	93.43%	91.10%	91.39%	94.13%	94.08%	92.18%	92.63%
ASR accuracy	49.83%	54.66%	46.11%	66.13%	69.52%	84.86%	66.00%	79.68%
Adaptive accuracy	93.17%	92.29%	89.85%	90.18%	92.98%	93.39%	90.92%	91.97%

ASR phones was permissible based on the rules in Table III, the ASR phone was encoded; otherwise, the G2P transcription was retained. In this manner, we were able to generate a third, allophone-adapted phonetic transcription of each utterance.

VI. EVALUATION OF SYSTEM PERFORMANCE

A. Transcription accuracy

The accuracy of each APT phase was evaluated using the phone error rate (PER) metric, which is defined by the following expression:

$$\text{PER} = \frac{S + D + I}{N_T} \times 100\%$$

where N_T represents the total number of phones in the manual transcription, and the sum $S + D + I$ denotes the edit distance between the manual and automatic transcriptions. The same alignment algorithm described in Section V was used to determine the edit distance and PER, while the accuracy was calculated using the equation $\%accuracy = 100\% - PER$. Table IV and Fig. 4 show the results obtained by applying single-language and multilingual data to each of the three phases.

The G2P system yielded the best performance among the three APT methods at 91.10%–94.13% accuracy, over 6% better than the best reported human inter-transcriber accuracy of 85% in [8]. Meanwhile, the ASR system performed the poorest of the three, with only the 3-hour Tagalog and 7-hour multilingual models operating within human accuracy. The adaptive system also returned accuracies of at least 5% better than the maximum inter-transcriber rate across all datasets; however, these results were consistently lower than the G2P accuracy rates by 0.63%–1.26%. As the ASR system had subpar accuracy compared to the G2P system in all eight test cases, it may have introduced more errors than it corrected during the allophone substitution phase, and therefore lowered the overall accuracy of the adaptive system instead of improving it as desired.

Notably, while the G2P and adaptive systems did not show any significant differences in performance with the increase in data from Set 1 to Set 2, the ASR system consistently yielded higher accuracies (ranging from 4.83% to 20.02% improvement) using the expanded data set. The ASR system may therefore potentially reach human accuracy levels for Cebuano and Hiligaynon transcription by increasing the amount of training data provided. Preliminary testing of the ASR system also showed improvements in accuracy by increasing the order of the n-gram language model, with accuracies of over 90% achieved at $n \geq 5$ for the 3-hour Tagalog dataset and at $n \geq 6$ for the 7-hour multilingual dataset. Further study on the relationships between dataset duration, language model order, and ASR performance is recommended.

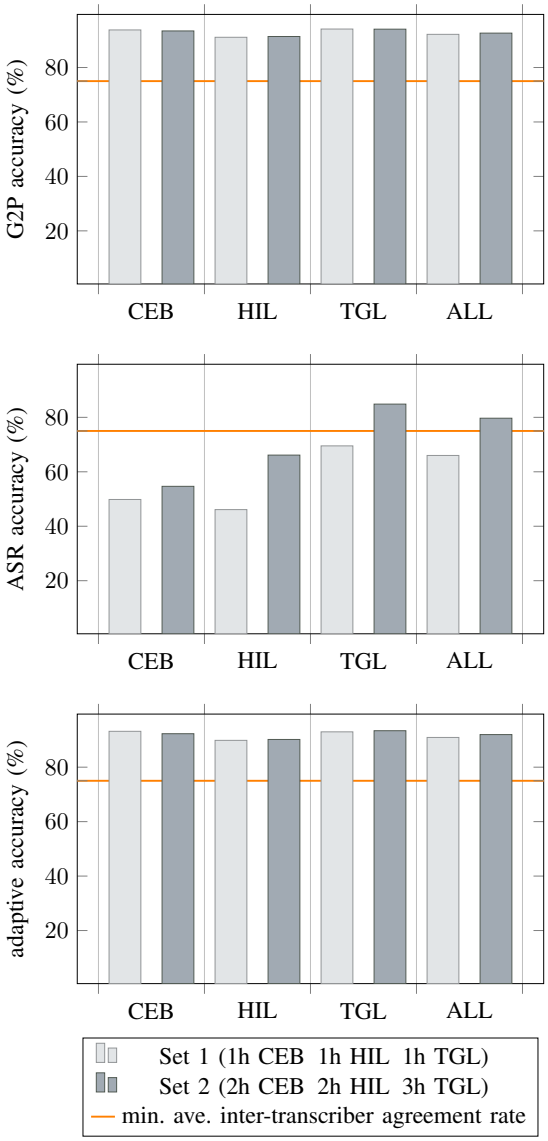

Fig. 4. Percent accuracies of APT methods versus manual transcriptions

B. Running time comparison

The running times of each system for different durations of input data were averaged and recorded as shown in Table

TABLE V
RUNNING TIMES WITH VARYING DATASET DURATIONS

Duration	G2P runtime	ASR runtime	Adaptive runtime
1h	2.86s	19m 30s	19m 34s
2h	10.79s	15m 54s	16m 5s
3h	14.85s	23m 49s	24m 6s
7h	13.55s	40m 13s	40m 31s

V. Of the three APT methods, the G2P system was observed to have the fastest transcription speeds (1.936s to 5.395s per hour of data), while the adaptive system was the slowest (5m 47s to 19m 34s per hour of data), having been implemented as a combination of both prior systems. Across the eight tests performed, all three systems exhibited a decrease in required transcription time of over 75% versus reported typical rates of human transcribers (180m per hour of data).

VII. CONCLUSION

In this paper, the implementations of three automated phonetic transcription methods (grapheme-to-phoneme conversion, automatic speech recognition, and adaptive substitution) are presented as alternatives to manual transcription of speech data in Tagalog, Cebuano, and Hiligaynon. As available speech corpora for Philippine languages are limited in number and scale, we trained and tested each system on small amounts of data in each of the target languages, ranging from 1 to 7 hours of recorded speech, and measured the speed and accuracy of the systems across varying lengths of input data.

Based on our results, the G2P system is evaluated as the most suitable alternative to manual phonetic transcription, having been shown to exceed human accuracy while reducing overall transcription time by over 98%, and is recommended for use whenever word-level transcriptions of the speech data to be processed are available. In cases where only recorded audio of speech utterances is provided, the ASR models for Tagalog and mixed speech are also shown to perform at par with human accuracy while yielding faster transcription speeds, and may be further investigated for possible improvement in performance with increased amounts of training data and varying language model order. These systems serve as a competent baseline for future developments in APT for Philippine languages, and are expected to significantly reduce the costs of transcription, facilitating further research and advancements in Philippine linguistics and speech technology.

ACKNOWLEDGMENTS

We would like to thank Mr. Michael Manahan, Prof. Mary Ann Bacolod, and Ms. Theresa Iana Tan of the UP Department of Linguistics for their invaluable contributions in verifying our datasets. We would also like to thank our reviewers for their insightful comments, and Mr. Michael Bayona for his ever-present support.

REFERENCES

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig (eds.), *Ethnologue: Languages of the World*, 22nd ed. Dallas, TX: SIL International, 2019. Available: http://www.ethnologue.com.

[2] C. Lopes and F. Perdigão, "Phone recognition on the TIMIT database," in *Speech Technologies* (I. Ipsic, ed.), InTech, 2011.

[3] K. Lee and H. Hon. "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1642-1648, November 1989.

[4] M. Ravanelli, P. Brakel, M. Omologo, Y. Bengio, F.B. Kessler, "Improving Speech Recognition by Revising Gated Recurrent Units," in *INTERSPEECH 2017*, August 2017.

[5] J. Vanek, J. Michalek, and J. Psutka, "Recurrent DNNs and its Ensembles on the TIMIT Phone Recognition Task," in *20th International Conference on Speech and Computer*, September 2018.

[6] R. G. Sagum, R. A. Ensomo, E. M. Tan, and R. C. L. Guevara, "Phoneme alignment of Filipino speech corpus, in TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region, IEEE, October 2003.

[7] I. J. Chua and N. Eustaquio, "Automatic isolated word recognition system and phoneme-level segmentation of the Filipino speech corpus using HTK. Undergraduate project, Department of Electrical and Electronics Engineering, University of the Philippines, 2006.

[8] C. Cucchiarini and H. Strik, "Automatic phonetic transcription: an overview," in *15th International Congress of Phonetic Sciences*, August 2003.

[9] N. Barroso, K. L. de Ipina, and P. M. Calvo, "An automatic and adaptive phonetic transcriber for the Basque language," in *4th International Work Conference on Bio-inspired Intelligence*, IEEE, June 2015.

[10] B. Schuppler, S. Grill, A. Menrath, and J. A. Morales-Cordovilla, "Automatic phonetic transcription in two steps: Forced alignment and burst detection," in *Statistical Language and Speech Processing: Second International Conference*, October 2014.

[11] E. P. Wolfenden, *Hiligaynon Reference Grammar*. Honolulu: University of Hawaii Press, 1971.

[12] P. Schachter and F. T. Otanes, *Tagalog Reference Grammar*. Berkeley and Los Angeles, California: University of California Press, 1972.

[13] A. A. Bollas, "Comparative analysis on the phonology of Tagalog, Cebuano, and Itawis, tech. rep., University of the Philippines Diliman, August 2013. Available: http://www.academia.edu/4427395.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[15] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, 1981.