Philippine Languages Database: A Multilingual Speech Corpora for Developing Systems for Philippine Spoken Languages

Rhandley D. Cajote, Rowena Cristina L. Guevara
Michael Gringo Angelo R. Bayona, Crisron Rudolf G. Lucas
University of the Philippines Diliman, Philippines
Trinity College Dublin, Ireland
University College Dublin, Ireland
{rhandley.cajote, rowena.guevara}@eee.upd.edu.ph
bayonam@tcd.ie, crisron.lucas@ucdconnect.ie

Abstract

Previous efforts to collect Filipino speech were done in the development of Filipino-Speech Corpus, TAGCO, and Filipino-Bisaya speech corpus. These corpora, however, are either domain-specific, non-parallel, non-multilingual or relatively insufficient for the development of state-of-the-art Automatic Speech Recognizers (ASR) and Text-To-Speech Systems (TTS) which usually requires hundreds of hours of speech data. This paper presents the Philippine Language Database (PLD) - a multilingual corpora for the Philippine languages namely: Filipino, English, Cebuano, Kapampangan, Hiligaynon, Ilokano, Bikolano, Waray, and Tausug. PLD includes over 454 hours of recordings from speakers of the ten languages, covering multiple domains in news, medical, education, tourism and spontaneous speech. The applicability of the corpus has also been demonstrated in adult and children ASR, phoneme transcriber, voice conversion, and TTS applications.

Keywords: speech corpora, low-resource languages, Philippine languages

1. Introduction

The Philippines, being an archipelago subdivided into seventeen regions, is a home of more than 100 native languages. Based on a 2020 survey (Philippine Statistics Authority), the major languages include Tagalog[1] (39.9%), Bisaya (22.5%), Hiligaynon (7.3%), Ilokano (7.1%), Bikolano (3.9%), Waray (2.6%), Kapampangan (2.4%), Maguindanao (1.4%), Pangasinan (1.3%), Tausug (1%) and Maranao (1%). Tagalog, even though it is the mostly used language in the country, is still considered as low-resource language (Cruz and Cheng, 2020). There are efforts to collect spoken data like TAGCO (Mesa, 2020), Filipino and Bisaya Speech Corpus (Pascual et al., 2023), Filipino Speech Corpus (Guevara et al., 2002), and (Liao et al., 2019). A recent paper on Wav2Vec2.0 XLS-R also mentioned a Tagalog dataset included in BABEL dataset[2] (Babu et al., 2022). The Common Voice dataset by Mozilla also has ongoing data collection and preparation for the Tagalog language (Juma, 2021).

A detailed summary of these corpora in terms of size and domain can be seen in Table 1. From the table, it can be seen that these corpora are either domain-specific, non-parallel, and non-multilingual. The largest among the list is the Babel dataset comprising mostly of telephone conversations sampled at 8000 Hz. Filipino Speech Corpus from UP Digital Signal Processing Laboratory is the next largest but with only 75 hours of Filipino speech. Because of the limitations of these corpora, the development of speech technologies for the Philippine languages have been very slow as compared to the other languages like English, German, and French.

Thus, the Philippine Language Database under the Interdisciplinary Signal Processing for Pinoys (ISIP) project was funded by the Department of Science and Technology (DOST) to be a prime mover in the development of speech technologies for the Philippine languages. The mission of the project is to spur the growth of many language and education research endeavors in the country, igniting exciting new areas of research. The possible applications envisioned for this project include: (1) vocabulary reading lists with accompanying audio guides, (2) pronunciation and grammar tutors (through a grading device or in the form of a game). (3) virtual learning environments, web-based language exchange applications, language portals. (4) multimedia development. (5) computer-based applications relating to automatic speech recognition, speech synthesis (text-to-speech systems), and machine translation.

On the linguistics side, there are many related corpus linguistics activities that would benefit from

[1] Filipino is the national language and it is based primarily on Tagalog that is linguistically classified as an Austronesian or Malayo-Polynesian language (Guevara et al., 2002).

[2] The Tagalog dataset from the Babel program is made available by the Lingusitic Data Consortium and is available as a paid dataset Bishop et al. (2016)

Corpus	Languages	Type	Size
Filipino Speech Corpus (FSC) (Guevara et al., 2002)	Filipino	read and spontaneous speech	75 hrs
Filipino-Bisaya Speech Corpus (Pascual et al., 2023)	Filipino, Bisaya	read speech, medical domain	Filipino: 35.88 hrs, Bisaya: 31.85 hrs
TAGCO (Mesa, 2020)	Tagalog	read and spontaneous speech	4.27 hrs
Liao et al. (2019)	Bikol, Kapampangan	read and spontaneous speech	Bikol: 2.5 hrs, Kapampangan: 4.5 hrs
IARPA Babel, cited in Babu et al. (2022)	Tagalog	spontaneous, telephone speech	213 hrs
Philippine Languages Database	Bikolano, Cebuano, English, Filipino, Hiligaynon, Ilokano, Kapampangan, Pangasinan, Tausug, Waray-Waray	read and spontaneous speech	454.83 hrs

Table 1: Existing speech corpora for Philippine languages.

the corpora, such as (1) corpus-based lexicography, (2) phonetic data analysis, (3) preparation and delivery of corpus-based educational materials, (4) content analysis, (5) stylistics, (6) statistical studies, and (7) language heritage documentation.

2. Data Design and Collection

2.1. Design

The corpora is envisioned to serve as seed data in the development of various spoken language processing systems for different Philippine languages. We aimed to build a multilingual corpus comprised of ten (10) languages in the Philippines. These languages are Filipino, Cebuano, Hiligaynon, Ilokano, Bikolano, Waray, Kapampangan, Pangasinense, Tausug, and English (with Filipino speakers as L2 speakers). Each language considered in this corpus has read and spontaneous speech data collected from speakers from various regions, ages and gender. Common criteria were taken into consideration including high quality recording of spoken and read speech, representativeness of the language, inclusion of all relevant acoustic realizations of the basic sound unit used, wide textual coverage, and wide prosodic and speaking style coverage.

The recording prompts for Filipino speech data collection were first determined. Prompts for the read speech part were collected from different sources such as literary works and news articles, and also included texts that reflect daily and situational conversations. These prompts were either downloaded from publicly available sources in the Internet or used with permission from the publishers. The news articles specifically is a subset of a dataset used in a previous project on a culturomic analysis based on Filipino written news articles (Ilao et al., 2011). The prompts for the read speech part were designed such that reading it will not take more than one minute. For the spontaneous speech data collection, questions were written such that any speaker can answer them with ease and can talk extensively about the topic covered. Similarly, responses for questions in the spontaneous speech part is not allowed to exceed one minute. For the other Philippine languages, the Filipino prompts were translated by hired native language speakers so that we will have parallel data for the ten (10) languages.

2.2. Recording Setup and Process

The collection of speech data was done either in the research laboratory or via fieldwork at various locations in the Philippines to facilitate the enlistment of participants from different ages, gender and regions. The research laboratory hosts a pseudo-anechoic chamber – a sealed booth that is approximately 2m x 3m. Wedge shaped acoustic absorbers are also padded around the walls, allowing for a clean recording with a noise floor rating of 20dBA. The recording equipment used includes a condenser microphone and two monitor headphones as shown in Figure 1. A duplicate screen

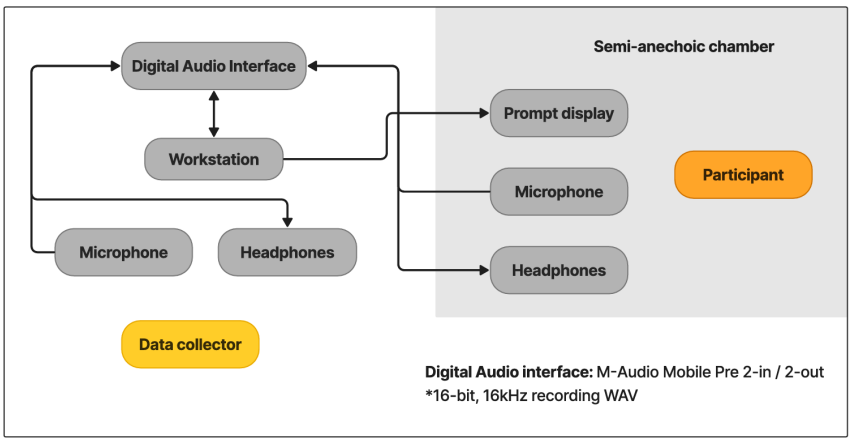

Figure 1: Diagram of the recording setup in UP Digital Signal Processing laboratory.

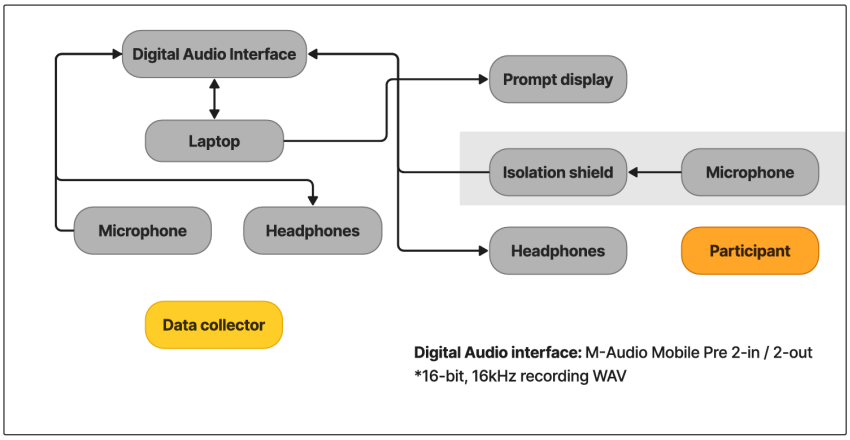

Figure 2: Diagram of the recording setup during fieldwork.

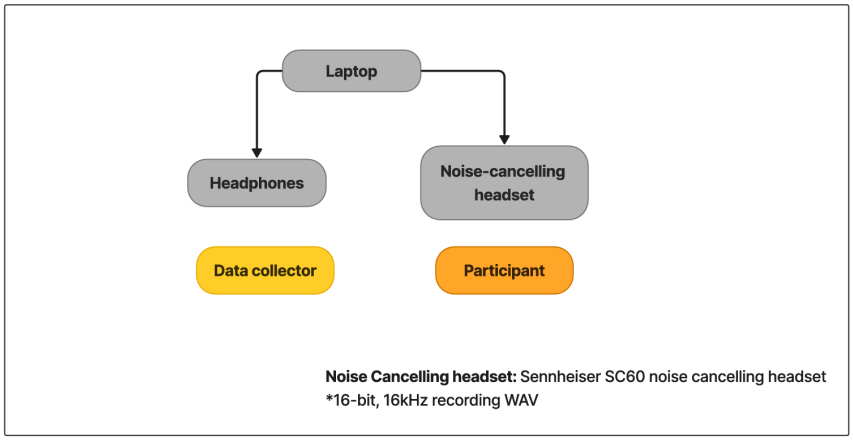

Figure 3: Variation of fieldwork recording setup using noise-cancelling headset.

was set up inside the chamber so that the recorded participant can see the prompts. A more portable setup was available during fieldwork, which used either a portable vocal booth and condenser microphones or noise-cancelling headsets as shown in Figures 2 and 3.

Participants enlisted in the data collection process were first informed about the details of the activity. Details include the project's funding information, the scope on the use of the recordings (for research purposes only), their rights regarding access and withdrawal of their recordings, and the anonymization of their personal data prior to the release of the corpus. Only when they agree to the terms of the activity will they be able to proceed with the recording. Participants expressed their

agreement by signing a participation agreement document.

A recording tool was used to facilitate the collection of speech data and speaker information. It is operated by a research assistant who will ensure that the speaker's information is correctly encoded, every prompt was correctly read, and all utterances were recorded. At the start of the recording session, details about the speaker is encoded via the recording tool. These include the speaker's age, gender, profession, first language and the first languages of the speaker's parents. The information about the first language is further differentiated by adding the region where the speaker or speaker's parents grew up, which is how we approximate the dialect spoken. The collected information is used to categorize the speakers and easily monitor the distribution of speakers per language according to age, gender and dialect.

After the speaker's information is encoded in the recording tool, the speaker is assigned a random subset of prompts to be read, presented one at a time on the screen. The selection of prompts is done automatically by the recording tool and prompts may be presented more than once in one recording session. The collection of read speech is immediately followed with the recording of spontaneous speech, where the speaker is given a random subset of questions to be answered. A recording session may include 200 to 400 read speech prompts and 1 to 3 questions for the spontaneous speech data collection. At the end of the session, the recording tool generates a log file containing the encoded speaker information, all the recorded prompts and corresponding filenames.

3. Corpora Details

3.1. Corpora Statistics and Current Status

Summary statistics for the PLD are shown in Table 2 where the information is broken down per language. The PLD currently contains over 340,000 recordings from over 1,000 speakers of 10 different Philippine languages. This corresponds to over 454 hours of recorded read and spontaneous speech, with an average utterance or audio length of around 4.7 seconds. Currently, a language corpus in the PLD has at least four hours of recordings (Tausug) to over 101 hours (Bikol). The combined recording prompts used for data collection correspond to over two million tokens, where a token can be a word, number, acronym, etc. used in the text, and does not include yet all the transcriptions for the spontaneous speech data collected as we are still in the process of transcribing this part of the corpora.

The distribution of speakers for each language according to age and gender is shown in Figure 4. For most languages, and regardless of gender, speaker ages cluster around 20 years old, as most of the participants are university students or young professionals. Exceptions are the age distributions for Hiligaynon (hil) and Kapampangan (pam), where speaker ages are more spread out, resulting into flatter and wider speaker age distributions.

The read speech part of the PLD corpora is already transcribed as the prompts are already matched with the corresponding correct recording. Meanwhile, the transcription of the spontaneous speech part by respective native speakers of the ten different languages is still in progress. Thus, the reported statistics on the total and unique tokens will change once all the spontaneous speech data have been transcribed.

3.2. Data Collection Timeline

The project started in July 2011 and ended in December 2014. During Year 1, from July 2011 to June 2012 the team has started to collect recordings in the lab for Filipino, Kapampangan and Pangasinense. Fieldwork recordings for Cebuano and Hiligaynon started in October 2011 and March 2012 respectively. In Year 2, from July 2012 to June 2013, recordings for Bikolano, Ilokano and Waray-Waray were added. Year 3, from July 2013 to December 2014 we started consolidating the data and continued to collect, when available, speakers for English. During this time we were able to contact a community of native Tausug speakers in Manila and solicited their help to facilitate recording this time in a laboratory recording set-up.

3.3. Corpora Structure

The corpora is organised as illustrated in Figure 5. Collected speech recordings for one Philippine language are stored in one directory, and are sorted according to speaker IDs, which currently are denoted by four-digit numbers. We split the IDs 0000 to 1999 among the 10 languages, having an initial ID allocation of 200 speaker IDs per language, but we will accommodate more speakers in any language, if there are any, and assign them speaker IDs from 2000 and above.

Each speaker ID folder contains the speech recordings, sampled at 16kHz and stored in WAV format. The transcripts for the read speech recordings are stored in a log file that is automatically generated by our recording tool after a completed recording session. For the spontaneous speech recordings, the recording tool uses the question displayed during the session as a placeholder transcript and is stored in the same log file, which is

Language	Gender	Speaker Count	Utterance Count	Audio Duration		Tokens	
				Total (h:m:s)	Average (s)	Total	Unique
Bikolano (bik)	F	121	39,260	60:55:58	5.5873	321,721	16,049
	M	85	27,684	40:17:36	5.2397	206,642	14,631
	all	206	66,944	101:13:35	5.4436	528,363	17,005
Cebuano (ceb)	F	86	34,956	35:47:01	3.6852	144,882	8,026
	M	66	27,477	27:44:58	3.6357	114,563	7,267
	all	152	62,433	63:31:59	3.6634	259,445	6,844
English (eng)	F	23	3,156	4:31:30	5.1617	29,376	4,363
	M	7	888	1:01:40	4.1675	8,050	1,483
	all	30	4,044	5:33:11	4.9434	37,426	4,729
Filipino (fil)	F	79	30,617	31:43:55	3.7311	205,346	10,861
	M	56	22,262	20:50:48	3.3712	138,088	7,994
	all	135	52,879	48:56:36	3.5796	343,434	11,481
Hiligaynon (hil)	F	48	17,079	21:49:43	4.6012	99,087	5,397
	M	43	14,908	19:21:58	4.6766	84,676	4,906
	all	91	31,987	41:11:42	4.6363	183,763	5,767
Ilokano (ilo)	F	64	15,429	25:46:37	6.0145	131,316	11,603
	M	60	14,513	25:44:43	6.3862	130,500	11,642
	all	124	29,942	51:31:20	6.1947	261,816	13,270
Kapampangan (pam)	F	104	35,024	49:42:02	5.1086	225,595	12,827
	M	83	26,926	40:37:22	5.4313	176,947	12,629
	all	187	61,950	90:19:25	5.2488	402,542	14,221
Pangasinan (pag)	F	12	3,959	6:01:36	5.4802	24,819	4,302
	M	6	1,945	3:07:00	5.7687	11,698	3,148
	all	18	5,904	9:08:36	5.5753	36,517	4,773
Tausug (tsg)	F	4	1,185	1:43:09	5.2236	12,684	2,023
	M	9	2,103	3:06:34	5.3233	7,279	1,536
	all	13	3,288	4:49:45	5.2874	19,963	2,376
Waray-Waray (war)	F	48	15,337	22:51:12	5.3643	94,764	6,071
	M	26	8,500	12:04:10	5.1118	52,704	5,518
	all	74	23,837	34:55:23	5.2743	147,468	6,291
Total	**-**	**1,030**	**343,208**	**454:49:43**	**4.7708**	**2,220,737**	**-**

Table 2: Summary statistics for the Philippine Languages Database. Below the each Philippine language name is its language ID in parenthesis, based from the ISO 639-3 standard, as published in Ethnologue (Eberhard et al., 2024) Note that the total token and unique token counts do not include yet the transcripts from the spontaneous speech part as this part of the corpora is still being transcribed.

then replaced by the actual transcript by hired transcribers.

The recording tool adopts a naming convention governed by the assigned speaker ID and the recording session date. The log file is denoted by two components, which follows the format <SPEAKER_ID>.<SESSION_ID>.log, where <SPEAKER_ID> is the assigned speaker ID and <SESSION_ID> is the session ID number. The session ID number is also composed of two components: the recording date and a random number generated by the recording tool to differentiate multiple recordings that were completed in the same day. In the example shown in Figure 5, we have speaker 0000 recorded on the 16th of August 2011 and assigned a random number 031856, giving us the log

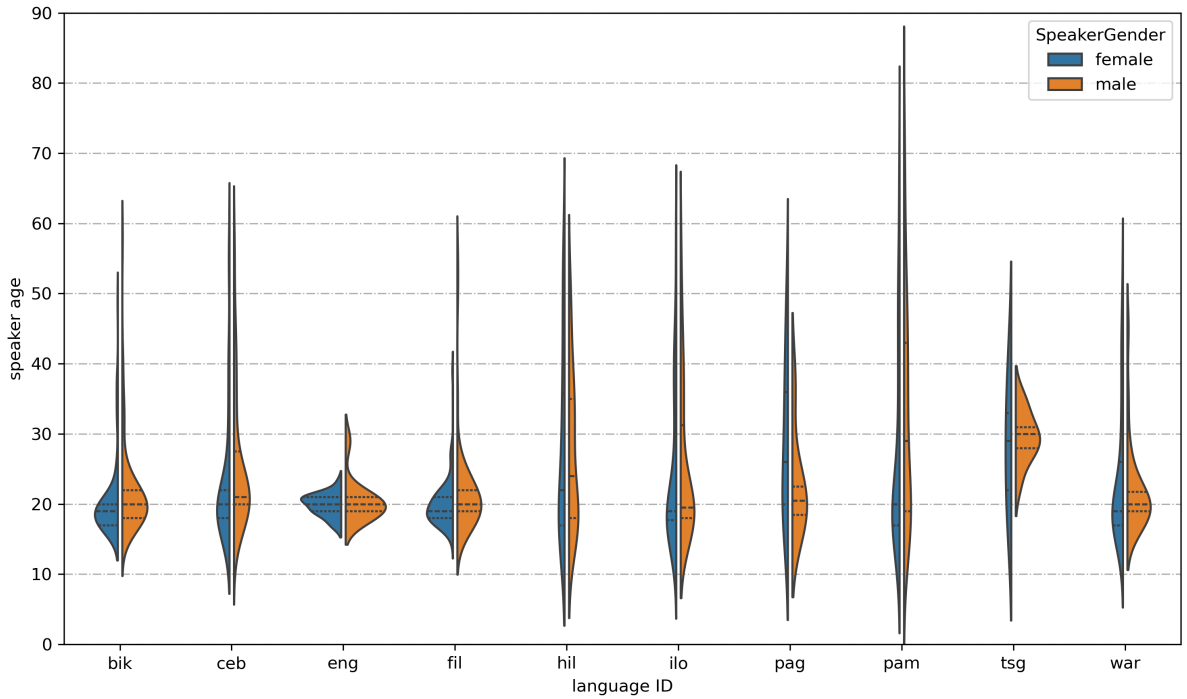

Figure 4: Age and gender distribution for each Philippine language included in the database. Language IDs used to label the violin plots are based from the ISO 639-3 standard, and the mappings to the corresponding Philippine language names are in Table 2.

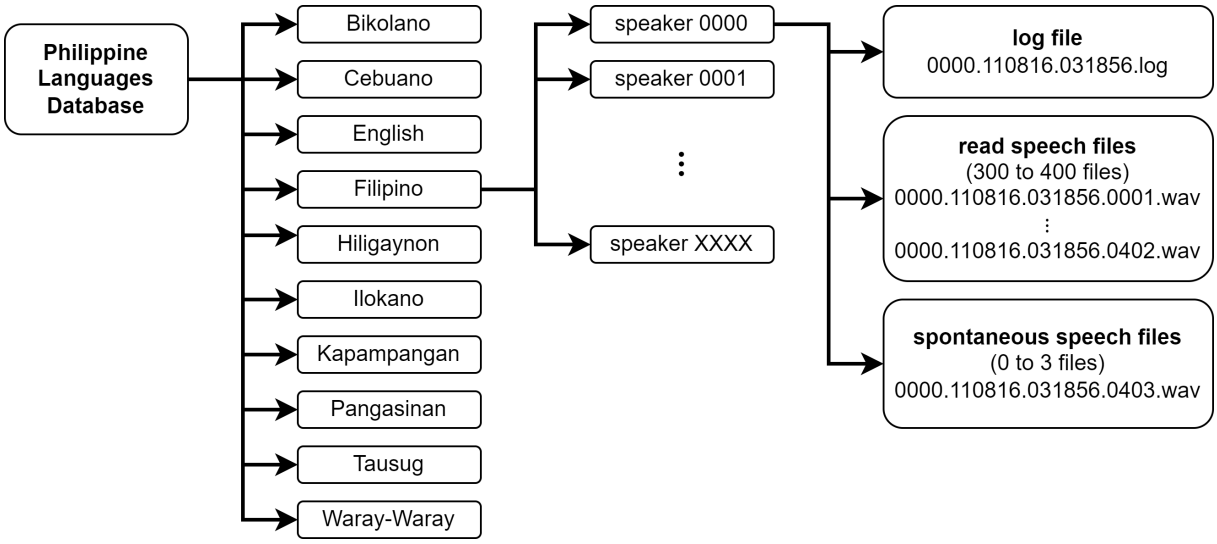

Figure 5: A diagram of the structure of the Philippine Languages Database. Collected recordings are grouped according to language and speaker, with each speaker corresponding to one folder. For each speaker, which in this example is speaker 0000, the corresponding folder contains the speech recordings stored in WAV format and the log file which contains the transcripts.

file name of 0000.110816.031856.log. Recorded utterances are stored following a similar format, with the addition of a fourth number denoting the order by which the utterance was recorded. Returning to our example, 0000.110816.031856.0001.wav is the first utterance recorded in the session.

3.4. Availability and Licensing

The PLD corpora can be accessed by filling out a letter of pledge indicating the purpose exclusively for research and academic use. A GitLab repository will be made publicly available that will include a sample of the data and the letter of pledge template which can be filled out by interested re-

searchers and emailed directly to the research laboratory (dsp@eee.upd.edu.ph). Upon creation, it is licensed under Creative Commons Attribution-NonCommercial (CC-by-NC 4.0).

4. Corpora Use

4.1. Speech-to-Text Systems

The Filipino corpus of the PLD was used by Ang et al. (2014) in developing a Filipino ASR which achieved 18.7% Word Error Rate (WER) on 2.8 hours of test data. The ASR implementation done was HMM-based with context dependency in the language model (LM), optimal feature space (OFS) training, and with Mel Frequency Cepstral Coefficients as features. In 2022, a study by Maranan (2022) also used the Filipino corpus for the development of a Filipino Children Speech recognizer (CSR) which is also HMM-based. Since the PLD corpus is adult speech corpus, Vocal Tract Length Normalization (VTLN) adaptation as well as pitch prosody-based augmentation was done to adapt to the CSR application. Maranan's system achieved 14.96 % WER for 40 minutes of test data.

Aside from ASR, a study by Aquino et al. (2019) used a subset of PLD in Filipino, Hiligaynon, and Cebuano for automatic phoneme transcription. In the study, the rule-based grapheme-to-phoneme (G2P) was compared to ASR-based method for phoneme recognition. In the study, G2P outperformed the ASR approach not only in terms of accuracy but also in runtime.

4.2. Speech Enhancement and Processing

A study by Gonzales et al. (2020) used the PLD subset of Filipino, Hiligaynon, Cebuano, and English for Voice-Conversion application. In his study, the parallel utterances were used for the target and source speakers for each language. He used wavelet modeling for the f0 contour along with the spectral parameters to improve the naturalness and overall quality of the voice-conversion. Using Mel-Cepstral Distortion (MCD) and Mean-Opinion Score (MOS) as evaluation metrics, the system was able to achieve 2.7 MOS for English (best) in terms of naturalness alongside the lowest F0:RMSE of 20.254. The system also performed better for intra-gender compared to inter-gender speaker voice conversion.

4.3. Text-to-Speech Systems

A study by Renovalles et al. (2021) used the 42,000 utterances of Filipino subset of PLD in the development of a Unit-Selection TTS system as well as the Tacotron2 TTS for Filipino. In the study, they also used voice conversion to augment the data by as much as 33,000 utterances. Overall, the Unit Selection performed better in the MOS test with 3.05 system level score. The Tacotron-2 with Data Augmentation only achieved 2.01 MOS.

5. Future Work

For future work, the developers of this corpus envisions the development of multiple low-resource speech applications extending beyond the developed Automatic Speech Recognition (ASRs), Speech Synthesis (SS) and Speech enhancement applications. Also, with the evolving research on natural and synthetic speech data augmentation, larger synthetic and hybrid corpora can be developed for pre-training large acoustics models (LAMs).

6. Acknowledgement

This study was supported in part by the project ICT for Education Digital Signal Processing for Pinoys (ISIP) Project 6: Philippine Language Database a Philippine government funded project thru the Department of Science and Technology - Grant in Aid (DOST-GIA) funds.

7. Bibliographical References

Federico Ang, Yoshikazu Miyanaga, Rowena Cristina Guevara, Rhandley Cajote, and Michael Gringo Angelo Bayona. 2014. Open domain continuous Filipino speech recognition with code-switching. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2301–2304.

Angelina Aquino, Joshua Lijandro Tsang, Crisron Rudolf Lucas, and Franz de Leon. 2019. OG2P and ASR techniques for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon. International Symposium on Multimedia and Communication Technology (ISMAC).

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2024. Ethnologue: Languages of the World.

Michael Gian Gonzales, Crisron Rudolf Lucas, Michael Gringo Angelo Bayona, and Franz De Leon. 2020. Voice Conversion of Philippine Spoken Languages using Deep Neural Networks. IEEE 8th Conference on Systems, Process and Control (ICSPC).

Rowena Cristina Guevara, Melvin Co, Evan Espina, Ian Dexter Garcia, Emerson Tan, Ryan Ensomo, and Ramil Sagum. 2002. Development of a Filipino speech corpus. In *3rd National ECE Conference*.

Joel Ilao, Rowena Cristina Guevara, Virgilio Llenaresas, Eilene Antoinette Narvaez, and Jovy Peregrino. 2011. Bantay-wika: towards a better understanding of the dynamics of Filipino culture and linguistic change. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 10–17, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hillary Juma. 2021. Introducing Common Voice Language Reps 2021/2022.

Edward Harold Liao, Kim Ganareal, Chirstian Clarence Paguia, Cesar Agreda, Manolito Octaviano, and Ramon Rodriguez. 2019. Towards the Development of Automatic Speech Recognition for Bikol and Kapampangan. In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5.

Jazzmin Maranan. 2022. An automated speech recognition system for phonological awareness of kindergarten students in filipino. 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).

Quennie Joy Mesa. 2020. TAGCO: A Tagalog Speech Corpus. volume 09. International Journal of Scientific & Technology Research (IJSTR).

Ronald Pascual, Judith Azcarraga, Charibeth Cheng, John Andrew Ing, Jian Wu, and Mark Louis Lim. 2023. Filipino and Bisaya Speech Corpus and Baseline Acoustic Models for Healthcare Chatbot ASR. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–5.

Philippine Statistics Authority. Tagalog is the Most Widely Spoken Language at Home (2020 Census of Population and Housing).

Edsel Jedd Renovalles, Crisron Rudolf Lucas, Franz de Leon, Angelina Aquino, and Izza Jalandoni. 2021. Text-to-Speech Systems for Filipino Using Unit Selection and Deep Learning. 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).

8. Language Resource References

Judith Bishop and Thomas Conners and Jonathan G. Fiscus and Breanna Gillies and Mary Harper and T. J. Hazen and Amy Jarrett and Willa Lin and María Encarnación Pérez Molina and Shawna Rafalko and Jessica Ray and Anton Rytting and Wade Shen and Evelyne Tzoukermann. 2016. *IARPA Babel Tagalog Language Pack IARPA-babel106-v0.2g LDC2016S13*. Linguistic Data Consortium, ISLRN 934-396-101-948-2.