Filipino-English Continuous Speech ASR: Towards Application of a Closed Captioning System for Philippine TV News Broadcast

Emmanuel M. Malaay[1*], and Rafael A. Ventura[1]
[1]*National University Philippines, Manila City*
**Corresponding Author: emmalaay@national-u.edu.ph*

Abstract: In this paper, the researchers present the development of a continuous Automatic Speech Recognition (ASR) System designed explicitly for Filipino newscasts, utilizing Convolutional Neural Networks (CNN). The ASR system is trained using a pre-existing 3-hour speech corpus. This corpus comprises speech utterances from various Filipino newscasts featuring news anchors and correspondents from multiple TV stations and news programs. The speech corpus contains a total of 10,346 words, encompassing English, Filipino, and Taglish languages. Each utterance in the corpus has been meticulously transcribed at both the word and phoneme levels. To ensure clarity and focus on the speech, background noise, and sound effects were removed during transcription. The primary application of this ASR system is to develop a closed captioning system for Philippine TV news broadcasts. This system aims to provide accurate and precise captions displayed on TV screens, enhancing accessibility for viewers. This initiative aligns with the Philippine Republic Act 10905 (RA 10905), passed by Congress and became law on July 21, 2016. RA 10905 mandates that all television stations and producers of television programs provide subtitles for their broadcasts. The researchers used Kaldi, an open-source speech recognition engine to develop our ASR system. Kaldi employs a Weighted Finite State Transducer (WFST) algorithm, which is well-suited for handling the complexities of speech recognition. ASR system, designed for English and Filipino languages, achieved a word error rate (WER) of 3.95%. The data used for training and testing the ASR system were sourced from Philippine TV news broadcasts on YouTube, an online video-sharing platform. During data preparation, the researchers extracted only the audio component of the videos, excluding any audio effects and background music, to maintain a clear focus on the speech content.

Keywords: ASR; speech; recognition; captioning; Kaldi

1. INTRODUCTION

As mandated by law under the Philippine Republic Act 10905, Philippine TV stations must have closed captions on their programs, which will be replayed or broadcast again for the next 24 hours. The law aims to assist the deaf or hard of hearing to comprehend and appreciate TV programs and motion pictures (Republic Act No. 10905, 2016). The transcription process for these TV programs will take some time, specifically the news broadcast, because it typically runs on air for about two hours. Even though news programs use scripts during broadcasts, there are times when the news anchor gives a follow-up question from the report, which will require an impromptu answer from the reporter. This situation is one constraint in which the transcription process will take time, which will be validated later to ensure the accuracy and precision of the caption.

In the Philippines, everyday conversations often blend Filipino and English, a mix known locally as "Taglish." This linguistic phenomenon presents a

unique challenge, as it involves seamlessly switching between two languages. Speakers frequently combine prefixes or suffixes from one language with words from the other and often use loanwords that have been culturally adapted.

Automatic Speech Recognition (ASR) automatically converts an input audio or speech signal into an output text. This text can be the final output or serve as input for Natural Language Processing (NLP). ASR systems are widely used in various applications such as voice commands, language learning, and captioning systems because they recognize speech and convert it to text (Levis & Suvorov, 2012). To achieve optimal performance, an ASR system must be trained with a high-quality speech database or corpus, as demonstrated in studies by Salido et al. (2018) and Yamashita et al. (2018).

Although ASR systems are valuable tools for speech recognition applications, they are sometimes overlooked due to their flaws. One primary concern is the reduced performance in real-time speech recognition applications. The performance of ASR systems heavily depends on the quality of the speech corpus used for training. Researchers face challenges in collecting and recording speech utterances to develop their own speech corpora, especially considering the numerous dialects in the Philippines. Properly and adequately trained ASR systems can become reliable for real-time applications, minimizing errors and enhancing speech recognition tasks (Newatia & Aggarwal, 2018).

The researchers utilized Kaldi, an open-source speech recognition engine, to develop our ASR system. Kaldi employs a Weighted Finite State Transducer (WFST) algorithm, which is well-suited for handling the complexities of speech recognition. Kaldi supports various acoustic and language models and has a strong community and extensive documentation. However, it can be complex to set up and requires a good understanding of speech recognition concepts, which may not be as user-friendly for beginners as some newer tools.

Kaldi remains a powerful option for those who need a highly customizable and flexible ASR system, while newer models like Whisper and wav2vec 2.0 offer impressive accuracy and ease of use. Whisper by OpenAI is highly accurate and robust, supporting multiple languages and tasks like transcription and translation, but it may lack some enterprise features. Facebook AI's wav2vec 2.0 excels in benchmarks and is effective for low-resource languages, though it requires significant computational power. Mozilla's DeepSpeech is easy to use and flexible but less accurate with long recordings. SpeechBrain is user-friendly and supports various speech tasks, but its newer status means it has a smaller community and fewer resources compared to more established tools like Kaldi.

This study aims to develop a continuous speech Automatic Speech Recognition (ASR) system, which will later be integrated into a closed captioning system for Philippine TV news broadcasts. This system will help Philippine TV stations provide accurate and precise program captions. As a test bed, the researchers gathered data from news broadcasts on YouTube, an online video-sharing platform. For this purpose, the researchers focused solely on the audio component of the videos to concentrate on the speech content. Additionally, the data were meticulously validated to remove extraneous information such as audio

effects and background music, ensuring a clear focus on the speech.

2. METHODOLOGY

2.1 Data Preparation, Pre-processing, and Validation

As a test bed, data were gathered from YouTube, an online video-sharing platform. The data collection was limited to news broadcasts, and only the audio portions of the videos were considered to focus on speech. Sound effects and background music were removed to emphasize the speech content.

The resulting database consists of continuous speech from news anchors and correspondents from different TV stations nationwide. The corpus lasts approximately eight hours, sampled at 44.1 kHz and later decimated to 16 kHz using the ASR toolkit to meet the system requirements. Some of the audio data included background noise, which was reduced using the sound editing software Audacity to ensure clarity and focus on the speech.

During live broadcasts, some reports are scripted, while others are impromptu. The word list or dictionary includes Filipino and English terms, proper and common names, acronyms, and expressions of emotions such as hesitations and ad-libs. Correct capitalization and punctuation were not initially considered but will be addressed in the later stages of this study.

Several ethical considerations must be considered when gathering and using YouTube data for an ASR system. Privacy and consent are paramount; even though YouTube videos are publicly available, the individuals in these videos may not have consented to their data being used for research. Researchers should anonymize data to protect the identities of individuals and avoid using content that could lead to personal identification. Transparency is also crucial; researchers should communicate how the data will be used and ensure their methods are ethically sound.

2.2 ASR Toolkit Preparations

The ASR toolkit used for conducting experiments in this study is Kaldi (Povey, 2015), an open-source tool for developing speech recognition systems. Kaldi employs a Weighted Finite State Transducer (WFST) algorithm, which is implemented in the C++ programming language. The WFST algorithm provides a common and natural representation for Hidden Markov Models (HMMs), which are context-dependent. This is achieved with a pronunciation dictionary, grammar, and alternative outputs (Mohri et al., 2002). The weighted FST optimizes time and space requirements by distributing paths for the statistical approach, making it ideal for designing and developing speech recognition systems. Kaldi needs to be installed and configured on a Linux operating system. In this study, Kaldi was installed on a Linux Ubuntu OS, running on a Virtual Machine with a 2-terabyte hard drive capacity, 16 gigabytes of memory, and an Intel Core i7 7th generation processor.

Kaldi offers several advantages for developing speech recognition systems. It is highly flexible and customizable, supporting various modeling approaches like HMMs, GMMs, DNNs, CNNs, and RNNs. Its core algorithms are

optimized in C++ for efficient training and decoding, and it leverages advanced math libraries for speed. Kaldi provides comprehensive recipes for building speech recognition systems, making replicating and adapting state-of-the-art results easier. As an open-source toolkit, it allows for free access, modification, and extension, fostering a large and active community. Kaldi can handle large datasets and complex models, supports distributed computing for scalability, and offers real-time processing with GPU acceleration. Its widespread adoption in academia and industry, including by major companies like Microsoft and Amazon, underscores its reliability and effectiveness.

While Kaldi is a powerful tool for speech recognition, it has some limitations. Its complexity and steep learning curve require a solid understanding of speech recognition concepts and proficiency in programming. The toolkit is resource-intensive, demanding significant computational power and memory, which can be a barrier for those with limited hardware. Setting up Kaldi can be challenging due to its dependency on a Linux environment and multiple configurations. Additionally, Kaldi has fewer pre-trained models than newer ASR frameworks, often necessitating users to train their models from scratch, which is time-consuming and data intensive. The decoding speed can be slower than some end-to-end ASR models, impacting real-time applications. Lastly, while Kaldi has a strong community, its documentation can be dense and technical, making it less accessible for beginners who may need to rely on community forums for troubleshooting.

2.3 Language and Acoustic Modelling

The ASR system developed in this study supports two languages: English and Filipino. A single model was created to recognize both languages. This approach is based on the study by Malaay et al. (2018), which developed a noise-robust speech recognition system for isolated digits applied through a telephone channel for a disaster participatory toolkit.

Word- and phoneme-level transcriptions were used to build the language and acoustic models, and a vocabulary of approximately 5,300 words in both English and Filipino was used. This preparation was done in compliance with the requirements of the ASR toolkit. The resulting dictionary is a reference for training the acoustic model to recognize both languages.

The phonemes used in this study are based on the ARPAbet, a phonetic transcription system representing the sounds in the speech data. This approach was also utilized in the studies by Malaay et al. (2018) and Pascual and Guevara (2012).

aa	ae	ah	ao	aw	ay
b	ch	d	dh	dy	eh
er	ey	f	g	hh	ih
iy	jh	k	l	m	n
ng	o	oh	ow	oy	p
r	s	sh	t	th	uh
uw	v	w	y	z	zh

Fig. 1. *Phoneme Set*

Figure 1 indicates all the phonemes used during the transcription of the data in the speech database, which was based on the work of the Advanced Research Projects Agency (ARPA), wherein they were able to form a list of phoneme alphabet as part of their Speech Understanding Research project during the 1970s which they later called ARPAbet. These phonetic transcription codes are widely used as a reference for linguistics applications and are proven to perform well in different speech recognition systems, such as in the works of Ang et al., (2011) and Malaay et al. (2017, 2018).

ARPAbet improves phoneme representation by providing a standardized set of phonetic transcription codes that represent the phonemes and allophones of General American English using distinct sequences of ASCII characters (Klautau, 2001). This system simplifies the representation of speech sounds, making it easier to process and analyze them computationally. Unlike the International Phonetic Alphabet (IPA), which uses a wide range of symbols, ARPAbet uses a more limited set of ASCII characters, which are easier to handle in digital systems. This makes ARPAbet particularly useful in speech recognition and synthesis applications, as it allows for consistent and clear transcription of speech sounds, facilitating the development and training of acoustic models.

The challenge was the variety of sounds in the Filipino language, where some words were read based on how they were spelled. Because of this, a combination of two sounds or diphones was formed, and this challenge was solved. Examples of this based on Table I are /oh/ and /dy/.

3. RESULTS AND DISCUSSION

After 17 iterations with approximately 12 hours of system training time, triphone 2 (Delta-delta) performs the best decoding scheme amongst all models generated by Kaldi. The optimized model can be determined by computing its Word Error Rate (WER), calculated using the formula in (1).

$$WER = \frac{S+I+D}{T} \qquad (1)$$

where:

$S = substitution$

$I = insertions$

D = *deletions*
T = *total number of words in the dictionary*

Substitutions occur when the system fails to recognize the exact word, meaning it would alter the word based on the language and acoustic model. The substituted word can sound alike or the same length and characteristic of the input phrase. Meanwhile, insertions occur when the system inserts a word that is not in the input phrase. Lastly, deletions occur when the system deletes a word on the input phrase. On the other hand, the system's accuracy can be determined from (1) by subtracting the error from 100, as shown in (2).

$$Accuracy = 100 - \frac{S+I+D}{T} \quad (2)$$

Table 1. *Word Error Rate (WER) of each Acoustic Model*

Model	WER (%)	ccuracy (%)
Mono	9.78	90.22
Tri1	7.67	92.33
Tri2	5.53	94.47
Tri3	7.75	92.25

Based on the experiments conducted, triphone 2 acquired a WER of 5.53%, equivalent to a 94.47% accuracy. Summary of WERs from the other models are summarized in Table 1.

Table 2. *ASR Experiments*

EXPERIMENT	WER (%)	Accuracy (%)
Experiment No. 1	5.53	94.47
Experiment No. 2	4.50	95.50
Experiment No. 3	3.95	96.05

The data was randomized and divided into two sets: 80% for training and 20% for testing. After each experiment, the system was checked and fine-tuned to prevent overfitting. After three experiments, the study identified the optimized model for the ASR system, achieving a word error rate (WER) of 3.95%, which corresponds to an accuracy of 96.05%. The results of the experiments are presented in Table 2.

The results of this study were achieved through continuous tuning of the ASR system to optimize parameters such as the number of cepstral coefficients, frequency coefficients, and energy levels, all of which depend on the training corpus. After the third experiment, the differences in word error rate (WER) were minimal, indicating that the optimized model for the application had been determined.

Compared to existing literature, this study achieved better results in terms

of WER. One contributing factor is the training algorithm used. Previous studies by Ang et al. (2011, 2014a, 2014b) used 60 hours of data. They achieved an 18.7% error rate with a Hidden Markov Model (HMM) implemented through the Sphinx III ASR toolkit. In contrast, this study employed a Weighted Finite State Transducer (WFST) algorithm through Kaldi, which is sometimes referred to as a "reinforced" or "enhanced" HMM. Unlike traditional HMMs, WFST does not have distributed paths, resulting in broader calculations that, while time-consuming, contribute to more accurate results.

4. CONCLUSIONS

This study successfully developed an Automatic Speech Recognition (ASR) system that supports both Filipino and English, addressing the code-switching challenge common in conversations in the Philippines. The optimal model for a closed captioning application for news broadcasts was identified using Kaldi's open-source speech recognition engine. The system achieved a word error rate (WER) of 3.95%, equivalent to a 96.05% accuracy rate, indicating that the system was finely tuned and enhanced for this application and can be tested with live data.

A significant aspect of this study is the focus on language use in developing the ASR system. Previous work by Ang et al. (2014a, 2014b) established a baseline for code-switching applications but required further improvement due to a higher WER. In contrast, this study used data from conversations and reports that included both Filipino and English, resulting in better WER and accuracy.

The current dataset comprises approximately eight hours of speech. Expanding the duration of the database is the first step towards further improving the system. Future steps include incorporating background music and sound effects into the system training and capturing emotions (e.g., hesitations and ad-libs) and scene descriptions relevant to news reports. The database can also be expanded to cover a broader range of domains, reflecting the common practice of multilingual conversations in the country.

Additionally, the structure of phrases and sentences and correct capitalization and punctuation for proper and common names must be considered. Exploring other training algorithms and methods, such as machine learning and deep learning, could enhance the system's recognition and training capabilities.

Code-switching in Automatic Speech Recognition (ASR) presents several challenges. One major issue is the pronunciation variation that occurs when speakers switch between languages, which can significantly reduce the accuracy of ASR systems. Additionally, the need for more code-switching training data makes it difficult to develop robust models, especially for intra-sentential code-switching, where the switch happens within a single sentence. The complex grammatical structures and co-articulation effects in code-mixed utterances further complicate acoustic and language modeling. Moreover, there is often an unbalanced distribution of language usage, where one language might dominate, leading to biased models. These factors collectively make it challenging to design

ASR systems that effectively handle mixed-language input.

To further enhance the ASR system, the following steps involve expanding the dataset to include more hours of diverse speech, incorporating background elements like music and ambient noises, and capturing natural speech patterns such as emotions and ad-libs. Additionally, broadening the domain coverage to reflect the multilingual nature of conversations in the Philippines is crucial. Improving text processing for correct capitalization and punctuation, exploring advanced training algorithms like machine learning and deep learning, and addressing code-switching challenges are also essential. Finally, extensive real-time testing with live data will help evaluate and refine the system's performance, making it more accurate, reliable, and versatile.

To enhance the ASR system, exploring advanced algorithms such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) is essential. DNNs can model complex patterns, while CNNs are effective for processing spectrograms due to their ability to capture spatial hierarchies (Rella, 2023). RNNs, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), excel in modeling sequential data (Rella, 2023). Additionally, transformer models, like those used in OpenAI's Whisper, handle long-range dependencies and are highly effective for transcription and translation tasks (Seagraves, 2024). Self-supervised learning models, such as Facebook AI's wav2vec 2.0, leverage large amounts of unlabeled data, making them particularly useful for low-resource languages (Seagraves, 2024). Hybrid models, which combine different neural network types, can further improve ASR performance by leveraging the strengths of each model type (Parti, 2024). Exploring these algorithms can significantly enhance the accuracy and robustness of the ASR system.

5. ACKNOWLEDGMENTS

This work is partly supported by the National University Research and Innovation Office (NU RaIN) under Project No. COE-2018-1T-29, through the National University College of Engineering. We sincerely thank the project's student research assistants and consultants, Engr. Ronald John Cabatic and Engr. Michael Simora, for their invaluable assistance in completing this project. We also extend our acknowledgment to Engr. Armil Monsura served as the project leader. His contributions were invaluable, and we honor his memory.

6. REFERENCES

Ang, F., Burgos, M. C., & De Lara, M. (2011). Automatic speech recognition for closed-captioning of Filipino news broadcasts. In *Proceedings of 2011 7th International Conference on Natural Language Processing and Knowledge Engineering*, 328–333. https://doi.org/10.1109/NLPKE.2011.6138219

Ang, F., Guevara, R. C., Miyanaga, Y., Cajote, R., Ilao, J., Bayona, M. G. A., & Laguna, A. F. (2014a). Open domain continuous Filipino speech recognition: Challenges and baseline experiments. In *Proceedings of*

IEICE Transactions on Information and Systems, E97.D(9), 2443–2452. https://doi.org/10.1587/transinf.2013EDP7442

Ang, F., Miyanaga, Y., Guevara, R. C., Cajote, R., & Bayona, M. G. A. (2014b). Open domain continuous Filipino speech recognition with code-switching. In *Proceedings of 2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2301–2304. https://doi.org/10.1109/ISCAS.2014.6865631

Klautau, A. (2001). *ARPABET and the TIMIT alphabet*. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0066

Malaay, E., Simora, M., Cabatic, R. J., Oco, N., & Roxas, R. E. (2017). Development of a multilingual isolated digits speech corpus. In *Proceedings of 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 1–5. https://doi.org/10.1109/icsda.2017.8384452

Malaay, E., Simora, M., Cabatic, R. J., Oco, N., & Roxas, R. E. (2018). Noise-resistant telephone quality isolated digits ASR: Towards application in a disaster participatory toolkit. In *Proceedings of 21st Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA).*

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language, 16*(1), 69–88. https://doi.org/10.1006/csla.2001.0184

Newatia, S., & Aggarwal, R. K. (2018). Convolutional neural network for ASR. In *Proceedings of 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 638–642. https://doi.org/10.1109/ICECA.2018.8474688

Parti, A. (2024). *Automatic speech recognition – the ultimate guide.* https://pareto.ai/blog/automatic-speech-recognition

Pascual, R. M., & Guevara, R. C. L. (2012). Developing an automated reading tutor in Filipino for primary students. In *Proceedings of 2nd Philippine Conference Workshop on Mother Tongue-Based Multilingual Education (MTBMLE 2)*, Iloilo City: University of the Philippines Diliman.

Povey, D. (2015). *Kaldi-ASR* [Computer software]. http://www.kaldi-asr.com

Rella, S. (2023, June 12). *What is Automatic Speech Recognition? | NVIDIA Technical Blog*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/essential-guide-to-automatic-speech-recognition-technology/

Republic Act No. 10905. (2016). *An act requiring all franchise holders or operators of television stations and producers of television programs to broadcast or present their programs with closed captions option and for*

other purposes. Official Gazette of the Republic of the Philippines. https://www.officialgazette.gov.ph/2016/07/21/republic-act-no-10905/

Salido, J. A. A., Oco, N., Roxas, R., Malaay, E., Simora, M., & Cabatic, R. J. (2018). Isolated digit Filipino speech recognition through spectrogram image classification: Towards application in a disaster preparedness participatory toolkit. In *Proceedings of 2017 International Conference on Asian Language Processing (IALP), 31-35.* https://doi.org/10.1109/ialp.2017.8300539

Seagraves, A. (2024). Benchmarking top open-source speech models. Retrieved from https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Yamashita, R., Nishio, M., Gian, R. K., DO, & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

A Publication of National University
Research and Development Office