Towards End-to-End Training of Automatic Speech Recognition for Nigerian Pidgin

Amina Mardiyyah Rufai[†] arufai@aimsammi.org
African Institute for Mathematical Sciences, Dakar, Senegal

Afolabi Abeeb[†] aaabeeb@pgschool.lautech.edu.ng
Ladoke Akintola University of Technology, Ogbomoso, Nigeria

Esther Oduntan eoduntan@aimsammi.org
African Institute for Mathematical Sciences, Kigali, Rwanda

Tayo Arulogun otarulogun@lautech.edu.ng
Ladoke Akintola University of Technology, Ogbomoso, Nigeria

Oluwabukola Adegboro[§] gadegboro@aimsammi.org
African Institute for Mathematical Sciences, Kigali, Rwanda

Daniel Ajisafe[§] dajisafe@aimsammi.org
African Institute for Mathematical Sciences, Kigali, Rwanda

Abstract

The prevalence of automatic speech recognition (ASR) systems in spoken language applications has increased significantly in recent years. Notably, many African languages lack sufficient linguistic resources to support the robustness of these systems. This paper focuses on the development of an end-to-end speech recognition system customised for Nigerian Pidgin English. We investigated and evaluated different pretrained state-of-the-art architectures on a new dataset. Our empirical results demonstrate a notable performance of the variant Wav2Vec2 XLSR-53 on our dataset, achieving a word error rate (WER) of 29.6% on the test set, surpassing other architectures such as NeMo QuartzNet and Wav2Vec2.0 Base-100H in quantitative assessments. Additionally, we demonstrate that a pretrained state-of-the-art model does not work well out-of-the-box. We performed zero-shot evaluation using XLSR-English as the baseline, chosen for its similarity to Nigerian Pidgin. This yielded a higher WER of 73.7%. By adapting this architecture to nuances represented in our dataset, we reduce error by 59.84%. Our dataset comprises 4,277 recorded utterances from native speakers, partitioned into training, validation, and test sets. This study underscores the potential for improving ASR systems for under-resourced languages like Nigerian Pidgin English, contributing to greater inclusion in speech technology applications. We publicly release our unique parallel dataset (speech-to-text) on Nigerian Pidgin, as well as the model weights on Hugging Face. Our project and code are made available[*] to foster future research from the community.

Keywords: Automatic Speech Recognition, ASR, Nigerian Pidgin English, End-to-End

∗. https://amina-mardiyyah.github.io/asr-nigerian-pidgin/

†. Equal contribution.

§. Co-supervisory role.

1. Introduction

Automatic Speech Recognition (ASR) plays a crucial role in transcribing spoken languages into text. Although ASR systems have become prevalent for widely spoken languages (Tamgno et al., 2012), the application of these systems to African languages remains a challenge. Africa, a continent known for its linguistic diversity with over 2000 languages (Abbott and Martinus, 2019), has a rich linguistic history shaped by language contact, expansion, trade language development, changes and instances of language dearth (Epstein and Kole, 1998). Most of these languages have at least one million speakers each, contributing significantly to the global linguistic landscape (Epstein and Kole, 1998). Despite the linguistic richness, African languages face substantial challenges in terms of resources (Abbott and Martinus, 2019). The majority of these languages are low-resourced (Laleye et al., 2016), hindering the development and robustness of ASR systems, especially when compared to more extensively resourced Western languages like English, French, and German. This study focuses on 'Pidgin English', one of the most widely spoken languages in Africa, with an estimated 75 million speakers in Nigeria and 5 million in Ghana. Although there are variations of this language, our research concentrates on the "Nigerian Pidgin English", the most prevalent form in West Africa (Ogueji and Ahia, 2019).

Prior work created the first monolingual Pidgin text-to-text corpus and trained the first word vectors on this language (Ogueji and Ahia, 2019). This involved aligning these vectors with English word vectors to produce cross-lingual embeddings. Subsequently, improvements were made by using the English text as a pivot language in the target domain, improving the fluency and relevance of the Pidgin text (Chang et al., 2020).

Under the speech-to-text category, Blachon et al. (2016) began the first effort to develop human language technology (HLT) tools, specifically speech resources for Nigerian Pidgin. Their work focused on developing a speech corpus for a tokenizer, and an automatic speech system for predicting the pronunciation of words and their segmentation. However, these resources were only integrated into a software tool, and we are not aware of any discoverable platform where the data has been made readily available for research purposes.

In an effort to strengthen the Nigerian Pidgin language and its area of research, we compiled the first publicly available speech-to-text dataset on Nigerian Pidgin. We fine-tuned Nemo and variants of the Wav2Vec2.0 architecture on this custom dataset. We compared the outcomes of these models against the baseline and achieved a lower error rate of 29.6% on the test set. Our key contributions are;

1. To provide a publicly accessible (end-to-end) ASR system for Nigerian Pidgin

2. Provide a free speech corpus for Nigerian Pidgin, and

3. Present the first parallel (speech-to-text) data on the language as a benchmark for further research.

The following sections are structured into the methodology, results, discussion, and conclusion section. The methodology section details the use of the LIG-Aikuma app for linguistic documentation, data sources and speech recording with corresponding preprocessing steps. Under the Model Architecture section, we first explained the baseline model – Wav2Vec2 XLSR-53-English, and the adapted models – NeMo QuartzNet, Wav2Vec2

Base-100H, and Wav2Vec2 XLSR-53 (large) respectively − describing the design, training, and adaptations. The results section presents the metric used and compares the architecture performances. Finally, the conclusion summarises the findings, emphasises contributions, and suggests future research directions for ASR in low-resource languages.

2. Methodology

Developing an automatic speech recognition system requires an adequate amount of speech recordings and corresponding text data. However, curating these paired resources is challenging, especially for low-resource languages. This is because they are not readily or publicly available online. In this section, we describe the methodology used to collect speech recordings and textual data for building our automatic recognition system.

2.1. Textual Data

The traditional method for amassing substantial textual data for ASR systems involve sourcing texts from online platforms. While many languages benefit from readily available resources like Wikipedia corpora, Pidgin English presents a different scenario. Though widely spoken by millions of West Africans, it lacks adequate NLP resources (Ogueji and Ahia, 2019). We build on prior data efforts made on the text-to-text parallel corpus, which was crawled from various news websites. The total crawled data comprised 56,695 sentences and 32,925 distinct words, covering topics ranging from sports, politics, and entertainment to everyday life. From this, we selected a subset of 4,288 utterances for recording speech data, with each utterance averaging between 8 and 17 words. The average sentence length in the corpus is 86 characters, with a corresponding mean audio duration of approximately 17 seconds.

To understand the most dominant themes in the dataset, we performed a multi-stage unsupervised topic modelling analysis using BERTopic (Grootendorst, 2022). We start with an initial clustering, which was subsequently improved using KeyBert (Grootendorst, 2020), and Maximal Marginal Relevance (MMR). This reduced topic outliers and enhanced coherence. By re-assigning and merging similar topics into one, we come down to 15 distinct themes, where "General/Everyday Conversation," and "Government/Politics" emerged as dominant themes, as shown in Figure 1.

2.2. Speech Corpus

With the acquired text information, we proceed to collecting speech recordings to build the recognition system. Given the absence of an existing speech corpus, we carried out the task of recording and collecting our own speech data. We used the LIG-Aikuma app for recording (Gauthier et al., 2016). The software made extensive data gathering possible due to its interoperability with Android devices and easy-to-share feature. In total, we amassed 4,288 instances of speech data, captured from 10 native pidgin speakers (comprising of 5 males and 5 females). Their ages ranged between 20 and 28 years, and the data was captured in an environment with mostly minimal background noise, as much as possible.

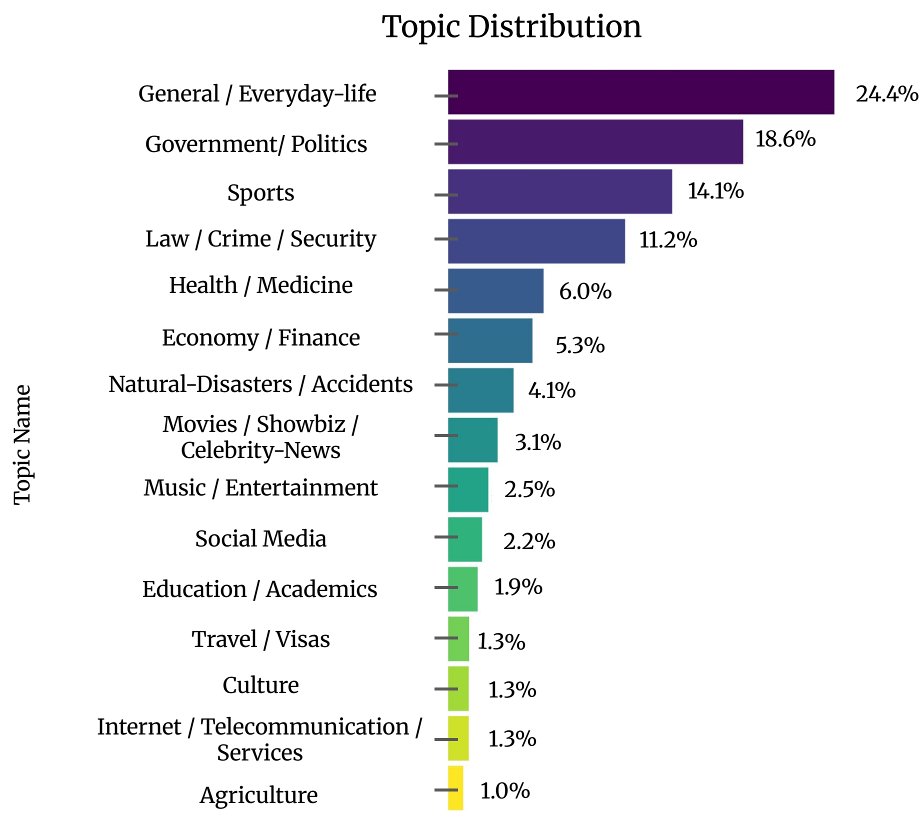

Figure 1: This shows the topic distribution in the Nigerian Pidgin text dataset, revealing dominant themes such as "General/Every-life" (24.4%), "Government/Politics" (18.6%), and "Sports" (14.1%), with -dominant themes in areas such as "Telecommunication" (1.3%) and "Agriculture" (1%).

2.2.1. Speech Data Preprocessing

Efficiently preprocessing speech data is crucial in developing robust speech recognition systems (Keerio et al., 2009). Therefore, our approach employs several key steps.

Audio Segmentation: To align with the requirements of the model architectures used in our experiments, we ensured that each audio file had a duration of no more than 30 seconds per chunk. Most of the audio files ranged between 1 and 17 seconds, with shorter audio clips typical in this dataset. To further standardise the data, all audio files were resampled to a 16 kHz sampling rate, ensuring consistency across the dataset and compatibility with our processing pipeline.

Additionally, corrupt files such as empty recordings, unintelligible speech, or audio containing only background noise were manually identified and removed by the data validator. These issues likely resulted from initial sentence boundary segmentation errors.

By resampling and excluding artifacts, our preprocessing pipeline ensured high-quality, consistent inputs, reducing the risk of errors during model training, while improving overall data integrity. After data preprocessing and filtering, the result is a final size of 4,277 speech recordings subsequently partitioned into training, validation, and test sets.

Feature Extraction: While core principles like resampling (16 kHz) and normalisation are common across architectures, specific feature extraction steps differ depending on the model. For NeMo QuartzNet, we applied and used Mel Spectrogram (Kriman et al., 2020) as input rather than a linear input. This extracts better feature representation such that differences in-between frequencies are more aligned to what humans perceive. In contrast, Wav2Vec2 models can handle raw waveforms directly − replacing traditional preprocessing steps with a convolutional feature encoder.

Data Augmentation: For NeMo QuartzNet, we utilised SpecAugment, a well-established technique for automatic speech recognition (Park et al., 2019) and applied this to the feature inputs. For the Wav2Vec2 variants, we followed the augmentation protocols provided in (Platen, 2021).

2.3. Model Architectures

We evaluated our data set on several state-of-the-art architectures. We briefly discuss key components of these architectures below:

2.3.1. Wav2Vec2 XLSR-53-English

To test the generalisation of a pre-trained state-of-the-art architecture on our custom Nigerian Pidgin dataset, we evaluate the Wav2Vec2-XLSR-53 English model (Grosman, 2021) − used as a zero-shot baseline due to the absence of a vocabulary dictionary required by available Wav2Vec2 models. This model was chosen for its linguistic closeness to Nigerian Pidgin. It was evaluated without fine-tuning to assess its generalisation to Pidgin English.

2.3.2. NeMo QuartzNet

The NeMo QuartzNet, as detailed in (Kriman et al., 2020), is based on a convolutional neural network framework and is trained using the CTC loss function (Graves et al., 2006) Notably, this model exhibits similarities to its forerunner, the NeMo Jasper architecture Li et al. (2019). A notable deviation lies in the adoption of 1D time-channel separable convolutions, diverging from the conventional 1D convolution employed in Jasper. The model was fine-tuned for 30 epochs using a greedy CTC decoder. We used NovoGrad as the optimizer with the same learning rate of $1e-4$ and weight decay of $1e-3$.

2.3.3. Wav2Vec2 Base-100H

Wav2Vec2 (Baevski et al., 2020) is an effective framework for learning powerful representations from speech data. Its self-supervised training methodology allows for seamless model pre-training on unlabelled data, which subsequently can be fine-tuned on small labelled data as a downstream task. We finetuned Wav2Vec 2.0 on Nigerian Pidgin and unfreeze the feature encoder to refine the model's ability to predict specific words or phonemes particular to the language. The model was trained for 30 epochs but with an AdamW optimizer (Loshchilov and Hutter, 2019). The learning rate is set to $1e-4$.

2.3.4. WAV2VEC2 XLSR-53 (LARGE)

The XLSR is built on the Wav2Vec 2.0 architecture, which learns cross-lingual speech representations from multiple languages. The model has a simultaneous learning process of quantized latent speech representations shared across languages (Conneau et al., 2020). These shared representations provide a strong generalisation capability for unseen languages. We used the same implementation setting as the base Wav2Vec 2.0 model.

3. Results and discussion

The word error rate (WER) is a key metric in automated speech recognition (ASR). It measures discrepancies between recognised and reference word sequences, as these sequences can sometimes be complicated by variable lengths. The word error rate can be computed as,

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{1}$$

Where S is the number of substitutions; D is the number of deletions; I is the number of insertions; C is the number of correct words; and N is the number of words in the reference ($N = S + D + C$).

The model with the lowest WER in the unseen test set is considered our best-performing model. Quantitatively, Wav2Vec2-XLSR-53 demonstrated the best WER on the unseen test set, highlighting its effectiveness in handling the nuances and variations of the language, courtesy of its robust representations. Table 1 below presents the full quantitative results, with Nemo baseline achieving the lowest WER at a score of 0.566 signalling the limitation of fully-supervised methods to cross-domain generalisation task. We also show qualitative results for the XLSR model in Table 2. Our method is able to capture nuances specific to the target language such as "sabi" (to understand), "tok" (talk), "pipo" (people) etc. − signalling a reasonable generalization ability. However, we noticed that it sometimes struggles with numbering e.g. "217" instead of "27" , as shown in Table 3.

Table 1: This table shows the quantitative results across all architectures used in this study. The best performing scores are indicated in **Bold**. The lower the WER, the better. Notably, Wav2Vec2 XLSR-53 achieved the lowest WER on validation and test sets.

Model	**Val WER**	**Test WER**
XLSR-53-English (Zero shot)[†]	–	0.737
NeMo QuartzNet	0.547	0.566
Wav2Vec2 Base-100H	0.397	0.373
Wav2Vec XLSR-53 (Ours)	**0.316**	**0.296**

Table 2: Qualitative Results for Wav2Vec2 XLSR-53, our best performing model. These demonstrate the model's ability to transcribe Nigerian Pidgin speech, highlighting its effectiveness in capturing language nuances.

Reference	**Our Prediction**	**Zero Shot Prediction**
• pipo and all di poor pipo wey govment gats take care of	pipo and all di poor pipo wey govrment gats take care of	people and ol the poor peopleway government gats take care of
• so dat one con mean say no show for dem next year	so dat one con mean say no show for dem next year	so thats on't calm me in senushu for them next year
• i no see why we no get proper health insurance	i no see why we no get proper health insurance	i' non's the why we know gepepiorts ensurance
• di social workers want make him pay her bill so dem	di social workers want make him pay habi so dem	in social wockarse one making paya bisodan

Table 3: Failure case for Wav2Vec2 XLSR-53. This highlights a limitation in transcribing Nigerian Pidgin speech, showing how numeric elements may be misinterpreted, thus affecting transcription accuracy.

Reference	**Prediction**
dem go wan kill chief femi fani kayode wife 27	dem go wan kill chief femi fani kayode wife 217

The quantitative results from our study on ASR systems for Nigerian Pidgin English reveal significant implications for both research and practical applications. The evaluation metric used, WER, demonstrates that Wav2Vec2 XLSR-53 surpassed other models in accuracy and performance on both the validation and the test set. We hypothesise the key implications and reasons for this superior performance to the following:

1. **Robustness and Adaptability:** Wav2Vec2 XLSR-53, built on the Wav2Vec 2.0 architecture, incorporates cross-lingual speech representations (XLSR), allowing it to learn latent speech features that are shared across multiple languages. This approach enhances the model's ability to handle variations and nuances specific to Nigerian Pidgin English, despite the language's under-resourced status.

2. **Fine-tuning Capability and Model Capacity:** The model's self-supervised pretraining methodology followed by fine-tuning on Nigerian Pidgin English speech data with increased model capacity plays a crucial role. The latter process refines the model's ability to predict specific phonemes and words relevant to the target language, resulting in improved accuracy compared to baseline models such as NeMo QuartzNet and earlier versions of Wav2Vec.

3. **Dataset Suitability:** Our study utilised a comprehensive dataset comprising 4,288 instances of speech data from native Nigerian Pidgin speakers. This dataset, collected

†. Validation WER for XLSR-53-English (Zero shot) is not reported, as this model was evaluated in a zero-shot setting on the test set only, without fine-tuning.

using the LIG-Aikuma app and augmented (during training) for robustness, provided ample training examples crucial for optimising Wav2Vec2 XLSR-53's performance.

The superior performance of Wav2Vec2 XLSR-53 in our study not only validates its efficacy but also highlights the potential of advanced ASR technologies in bridging the gap for under-resourced languages. This research contributes to the broader goal of inclusivity in artificial intelligence applications, particularly in diverse linguistic contexts such as those found in Africa.

4. Ethics and Limitation

Our data collection process adhered to ethical standards such as informed consent and speaker privacy. Each speaker was contacted via official email requesting their consent to publicly release the dataset. All participants provided explicit consent by replying affirmatively to the email. To ensure speaker anonymity, unique ID tags were assigned to the voice recordings, thereby concealing any personally identifiable information. While valuable as an initial resource, the dataset remains limited in size and speaker diversity, and it does not fully capture the dialectal variations across regions − limiting generalisation and robustness of ASR models trained on this resource. Increasing these variations would be beneficial for cross-domain performance.

5. Conclusion

In this study, we introduced parallel speech-to-text data for Nigerian Pidgin as a benchmark to support accessible and reproducible research in automatic speech recognition (ASR) systems. Using a zero-shot approach, the pretrained model produced a WER of 73.7%, reflecting the challenge of recognising Nigerian Pidgin without task-specific training. However, after fine-tuning on our curated dataset, the WER was substantially reduced to 29.6%, marking a significant advancement in ASR performance for this under-resourced language. This improvement demonstrates the value of domain-specific data and adaptation strategies in building effective speech technologies for African languages.

Future work should focus on enhancing the quality and diversity of the training and evaluation dataset. Broader demographic and geographic representation in speech samples can improve model generalisation, and techniques such as unsupervised test-time speaker adaptation may further reduce error rates. Continued collaborative research is essential to ensure that languages like Nigerian Pidgin are well-represented in the digital sphere, fostering inclusive and equitable language technologies.

Acknowledgments

The authors would like to thank generous volunteers who have contributed enormously to data collection, validation, and anonymous reviewers who have provided valuable feedback on the writing.

References

Jade Abbott and Laura Martinus. Benchmarking neural machine translation for southern african languages. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 98–101, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66, 2016.

Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. Unsupervised pidgin text generation by pivoting english data and self-training. *arXiv preprint arXiv:2003.08272*, 2020.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Edmund L. Epstein and Robert Kole. *The Language of African Literature*. Africa World Press, 1998.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*, 2016.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376, 2006.

Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL https://doi.org/10.5281/zenodo.4461265.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jonatas Grosman. Fine-tuned xlsr-53 large model for speech recognition in english. Hugging Face, 2021. URL https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin. On preprocessing of speech signals. *International Journal of Signal Processing*, 5(3):216–222, 2009.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE, 2020.

Fréjus A.A. Laleye, Laurent Besacier, Eugène C. Ezin, and Cina Motamed. First automatic fongbe continuous speech recognition system: Development of acoustic models and language models. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 477–482. IEEE, 2016.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kelechi Ogueji and Orevaoghene Ahia. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *arXiv preprint arXiv:1912.03444*, 2019.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Patrick Von Platen. Github repository for wav2vec code adaptation:. https://github.com/patrickvonplaten/notebooks, 2021.

James K. Tamgno, Etienne Barnard, Claude Lishou, and Morgan Richomme. Wolof speech recognition model of digits and limited-vocabulary based on hmm and toolkit. In *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, pages 389–395. IEEE, 2012.