Proceedings of the CLTW 4 @ LREC2022 , pages 104–109
Marseille, 20-25 June 2022

BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus

Stephen John Russell, Dewi Bryn Jones, Delyth Prys
Language Technologies Unit, School of Linguistics
Bangor University, Bangor, Wales, UK
{stephen.russell, d.b.jones, d.prys}@bangor.ac.uk

Abstract

This paper presents the design, collection and verification of a bilingual text-to-speech synthesis corpus for Welsh and English. The ever expanding voice collection currently contains almost 10 hours of recordings from a bilingual, phonetically balanced text corpus. The speakers consist of a professional voice actor and three amateur contributors, with male and female accents from north and south Wales. This corpus provides audio-text pairs for building and training high-quality bilingual Welsh-English neural based TTS systems. We describe the process by which we created a phonetically balanced prompt set and the challenges of attempting to collate such a dataset during the COVID-19 pandemic. Our initial findings in validating the corpus via the implementation of a state-of-the-art TTS models are presented. This corpus represents the first open-source Welsh language corpus large enough to capitalise on neural TTS architectures.

Keywords: text-to-speech, TTS, speech synthesis, speech corpus, open-source, bilingual, Welsh, English

1. Introduction

The prevalence of speech interfaces across modern society, often seen as "an essential component in many applications such as speech-enabled devices, navigation systems, and accessibility for the visually impaired" (Arık et al., 2017), poses an interesting challenge when developing text-to-speech solutions for bilingual communities. Tadmor (2009) noted that most languages contain loanwords from one or more other languages to some extent or another, however speakers in bilingual communities often take this further by alternating between languages mid sentence or word (Haspelmath and Tadmor, 2009). This linguistic trait, commonly referred to as "code switching" (Nilep, 2006), requires speakers to "include morphemes from two or more of the varieties of their linguistics repertoire" (Myers-Scotton, 2017). In order to ensure fair and unbiased access to technology in bilingual communities, and to help prevent against the threat of "Digital Language Extinction" (Rehm, 2014), it is essential that synthesised voices are equally proficient at articulating and disseminating the required information in both languages.

A member of the Celtic languages, Welsh has coexisted alongside English, in the United Kingdom, for hundreds of years (Cooper et al., 2019). Bilingual Welsh-English speakers often utilise code switching, by using English words mid sentence for named entities, convenience or to assist in communicating with learners. To address the phenomenon of code switching, previous works on bilingual Welsh-English text-to-speech synthesis have focused on statistical models, relying on "a bilingual pronunciation dictionary containing large numbers of words from both languages described phonetically with a series of phonemes" (Prys et al., 2021). Similar approaches can be seen for Mandarin and English (Chu et al., 2003; Zhiyong et al., 2009). More recent approaches to multilingual text-to-speech, exemplified by Casanova et al. (2021), have demonstrated how deep neural learning can be applied to multi-speaker datasets with impressive results. There are many benefits to a neural network based approach, synthesised voices demonstrate improved intelligibility and naturalness whilst also reducing the manual pre-processing and feature detection (Tan et al., 2021). However, "In comparison, deep neural models require substantially greater volumes of data than traditional TTS architectures" (Latorre et al., 2019), which can be prohibitive when working with lesser resourced languages.

In 2018, the Welsh Government released its Welsh Technology Action Plan, containing their plans for "technological developments to ensure that the Welsh language can be used in a wide variety of contexts, be that by using voice, keyboard or other means of human-computer interaction" Welsh Government (2018). Although previous works by Cooper et al. (2019) and Prys et al. (2021) have addressed many of the issues outlined, a comprehensive text-to-speech corpus, large enough to utilise advances in neural network architectures was not yet available.

In this paper, we present the first instalment of the Bangor University TTS Corpus,[1] a phonetically balanced, bilingual, Welsh-English corpus and prompt set, released under an open CC0 1.0 license.[2] The corpus contains 12,200 text prompts divided into

[1] https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-talentau-llais

[2] https://creativecommons.org/publicdomain/zero/1.0

9500 Welsh language prompts and 2700 in English along with voice recordings consisting of 2 male and 2 female native Welsh speakers, one of whom is a professional voice artist. Other lesser resourced languages have taken the approach of gathering source material from news reports or literature, such as the work done by Mussakhojayeva et al. (2022), and then post processing the recordings into shorter segments. We however, present the construction of a curated "phonetically balanced corpus" (Gibbon et al., 2012) and its purpose in providing a platform from which to build more specialised and domain specific voices.

Further to the prompt set and voice recordings, we also present our initial findings validating the corpus via a series of experimental implementations of a state-of-the-art TTS system, based on the VITS (Kim et al., 2021) architecture. The corpus presented is significantly smaller than intended and as such we detail the processes and adaptations implemented to deal with the disruptions and circumstance during the initial data collection phase. The synthesised voices were evaluated reading a selection of news articles in both Welsh and English to assess their intelligibility.

The organisation of this paper is as follows: Section 2 reviews a selection of Welsh language corpora. In Section 3, we describe the process of creating and compiling the data and give a statistical overview of the released corpus. A series of experiments and their architectures are set out in Section 4. Section 5 discusses the challenges faced implementing a bilingual Welsh-English text-to-speech solution and future research within this domain. This work is concluded in Section 6.

2. Related Work

Welsh is classified as a lesser resourced language, however "the availability of both text and speech corpora for Welsh has much improved in recent years" (Prys et al., 2022b). Cysill Ar-lein, the Bangor University free online spelling and grammar checker, has produced a corpus of over 400 million tokens by collecting user input, further reading can be found via Prys et al. (2022b). The CorCenCC corpus (Knight et al., 2021), contains in excess of 11 million tokens, annotated with parts of speech and semantic meaning. For the purposes of speech recognition, Mozilla's Common Voice is utilised extensively by over 1600 users, currently containing over 143 hours recordings. Open text-to-speech corpora, by comparison, are not so readily available, with the complete absence of an appropriately sized corpus for machine learning. The WISPR project (Prys et al., 2004) is one such corpus and contains 3 hours of speech recordings of a single speaker, with excerpts from the Bible and an undergraduate dissertation totalling 616 sentences of varying length. Off the back of this corpus one of the first Welsh text-to-speech voices was created, however it was restricted by the technology available, and as such produced only moderatly intelligible audio containing significant audio glitches. By 2016 this same dataset was utilised to create an open source voice, for the Welsh Digital Assistant Macsen (Jones, 2020), which by all accounts sounded much more natural, by utilising the MaryTTS framework (Schröder and Trouvain, 2003). The same technology is used for Lleisiwr, a project which sets out to create personal synthetic voices for users that may be at risk of losing their ability to speak. Prys et al. (2022b) provide further reading on the functionality of Lleisiwr.

Looking beyond the Welsh language, bilingual text-to-speech systems have been considered with a similar approach for the Manderin-English language pair (Chu et al., 2003; Zhiyong et al., 2009), also utilising bilingual pronunciation dictionaries to good effect. The presence of foreign words within a larger, machine learning ready corpus, has been considered by Mussakhojayeva et al. (2022) during the expansion and improvement of their KazakTTS corpus. The foreign words used here are limited to a subset of very important words imported from Russian which improves the ability to digitally communicate in Kazakh but falls short of a fully bilingual solution. Casanova et al. (2021) approaches a bilingual solution by utilising a combination of mono and multi speaker datasets such as Mozilla's Common Voice, LibriVox and LJSpeech to name but a few. The datasets are used to create pre-trained models that can be used to create one-shot voices via transfer learning mechanisms. This however produces a variety of voices from a single user due to the number of speakers used for training.

3. BU-TTS Corpus

This section details the creation of a phonetically balanced prompt set as well as the process undertaken to record it. We present both our intended methodology and the resulting processes that were required to complete the project.

3.1. Phonetically Balanced Text Corpus

The texts used to create our prompt set come from Mozilla's Common Voice, which in turn were curated from a variety of sources including self generated data from the Cysill Ar-lein corpus (Prys et al., 2016) and translations for under-represented categories such as recipes as well as open source or out of copyright external sources such as wikipedia, Twitter, Welsh language books and translations of selected English language books.

Due to the initial limitations of Common Voice, the sentences are limited in length to no more than 14

words. The process of segmenting and truncating longer texts into shorter sentences was undertaken by the terminologists and linguists within the team to ensure the quality and suitability of each sentence for open source distribution. Offensive or inappropriate language was removed with a focus instead placed on isolating interesting and easy to read sentences, appropriate for all ages. This process required a large amount of editing to segment longer text into sentences and remove any errors present in the text or to update old fashioned vocabulary, style or orthography. Further reading on the creation of this corpus will be available in the forthcoming Language and Technology in Wales: Volume II (Prys et al., 2022a).

From this master list a sub set of phonetically balanced prompts were compiled using the pre-built tools released in the MaryTTS toolset (Schröder and Trouvain, 2003), in conjunction with the Bangor University Pronunciation Dictionary (Prys and Jones, 2018). The pronunciation dictionary contains phonemes for both north and south Welsh accents, ensuring that the prompts chosen would represent a diverse range of dialect choices. Further to the phoneme coverage, the prompts were also checked to ensure values from the wordlists of the most common word-forms in Welsh, and the most common English words used in Welsh (Prys and Jones, 2019) were present in the final selections. The resulting 12,200 sentences form a series of 5 unique subsets, containing both Welsh and English prompts, each individually phonetically balanced to give an even distribution of phonemes.

3.2. Recording Process

Recording initially took place in the language laboratories at Bangor University. These laboratories are specially built to isolate recordings from outside sounds and, as such, provide an excellent low sound floor for recording as well as being fitted with a monitoring booth for supervising the recording process. In order to achieve an efficient and low noise recording process for amateur talents, iOS and Android apps, supporting the Sure MV88+ microphone, were developed for both the recording process and displaying prompts for the talents to read. In conjunction with the apps, an API web service and dashboard were constructed to collect the audio recordings and manage the users progress through the prompt set. Amateur talents were instructed to speak with a natural and relaxed tone whilst remembering to note punctuation and inflection as indicated in the sentences. Recordings were then reviewed via the dashboard and any non conforming recordings were discarded and re-introduced to the prompt set by the API service.

3.2.1. Remotely Recording Amateur Talents

We looked initially for amateur talents willing to donate their voices to the project and found many students at Bangor University eager to participate in the project. Having auditioned the voices we settled on 2 males voices with northern and southern accents and a female voice with a northern accent, whilst continuing to look for a 4th female voice with a southern accent. However due to the restrictions enforced by the COVID-19 pandemic, the laboratories were temporarily deemed unsuitable for data collection. Instead, we setup each of the voice talents with a mobile telephone and microphone to perform the recordings at home. Given the amateur status of the voice talents we quickly found that, without the direction of a supervisor, it was difficult to retain a consistent standard and rate of recording. At this point in the project, the potential benefits of such data to our voice cloning system Lleisiwr, where users are unlikely to have have professional recording equipment but are in great need of a voice, was highlighted. As such, we continued to gather as much data as was reasonable from the amateur talents, pursuing methods of audio cleaning and verification of the sparse and noisy recordings.

3.2.2. Professional Voice Talents

With an insufficient quantity of low quality data being produced by the amateur talents, we turned to a professional talent and recording company to complete the 4th voice. We provided them with the entire corpus of sentences, each tagged with an appropriate file name to be used for recordings. Once recorded the files were checked for accuracy and any silence was trimmed from the files. The professional talent was instructed to read the prompts in a neutral style, ensuring to emphasise where question marks and exclamation marks were present in the sentence. This process produced a plethora of high quality data in a relatively short time frame and forms the backbone of the BU-TTS corpus.

3.3. Corpus Overview

The format of the BU-TTS corpus is similar to that of the LJSpeech corpus (Ito and Johnson, 2017) where audio files are kept in a directory named "wavs", adjacent to a metadata CSV containing the file names of the wavs and the transcribed text. An illustration of the generic directory structure can be found in Figure 1. The recordings are released in 48 kHz 16-bit mono WAV files whilst the text is encoded with the UTF-8 format. All in, there are 9.8 hours of recordings from 4 contributors, the division of the speakers, and their number of recordings, can be found in Table 1 with the final language distribution of the prompt set outlined in Table 2.

4. Dataset Validation Experiments

To validate the potential of the corpus on neural network architectures, we made use of the Coqui-ai TTS repositories.[3] Coqui provide open-source frameworks

[3]https://github.com/coqui-ai/TTS

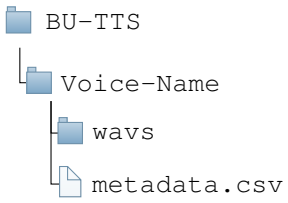

Figure 1: Generic Directory Structure.

Speaker	Batch	Recordings
F1	5	12,200
F2	2	3,653
M1	2	2,847
M2	2	2,630
Sum	N/A	21330

Table 1: Speaker Recording Distribution.

for both text-to-speech and speech-to-text that implement a variety of neural model architectures. Their libraries are intended for advanced text-to-speech generation and implement the latest research. Further to the codebase, Coqui-ai TTS is shipped with pre-trained text-to-speech models as well as tools for measuring dataset quality.

4.1. Experiment Architecture

Many text-to-speech models are based on a two stage architecture consisting of an initial aligner training phase and an independent vocoder training stage (Zeng et al., 2020; Ren et al., 2019). This can lead to long training sessions and requires the vocoder to be trained independently from the aligner. We instead chose to utilise the VITS (Kim et al., 2021) model architecture as it provides a simpler end-to-end process for speech synthesis. It was also decided to use exclusively graphemes to train the models due to multiple languages being used. This enabled us to focus on data curation and consolidation whilst also providing a benchmark standard from which to improve upon.

4.2. Single Speaker Experiments

Initial experiments were carried out using a single speaker VITS model with the southern accented female voice dataset, due to it being the only complete dataset and as such the only voice with a high enough volume of data for the neural architectures to be effective. This point was well illustrated when the largest dataset, with over 3000 recordings, from the amateur talents was utilised and only incomprehensible speech was produced. The successful model was trained using an NVIDIA RTX 3090 GPU for 3 days with the audio sampling kept at a full 44.1 KHz and 16 bit quality to attempt to retain the highest level of audio fidelity.

Lang	Batch	Prompts	Avg Words	Min Words	Max Words
cy	1	500	10	6	14
en	1	200	9	4	14
cy	2	1,500	10	4	14
en	2	625	9	4	14
cy	3	2,500	10	5	14
en	3	625	9	4	14
cy	4	2,500	10	5	14
en	4	625	10	6	14
cy	5	2,500	10	4	14
en	5	625	9	5	14
cy	All	9,500	10	4	14
en	All	2700	9	4	14

Table 2: Word - Sentence Distribution.

4.3. Transfer Learning & Multi-Speaker Experiments

Once a quality model had been achieved with a single speaker, attempts were made to use the lesser quantity and quality of data received from the amateur talents. Firstly we attempted using the pre-trained model for transfer learning and then subsequently via a multi-speaker implementation of the VITS model. During both experiments we utilised the cleaned and raw versions of the audio to get an understanding of any audio scrubbing requirements for future work.

4.4. Experiment Results

The trained text-to-speech models have only been informally tested, in house and at various live events, however initial reactions have been mostly positive and we can demonstrate an ability to code switch between Welsh and English within the same sentence. There are however still instances when words take the same form in both languages where errors will occur. Further to the ability to code switch, we have also demonstrated the ability to produce bilingual audio from large texts containing sequential Welsh and English content. Due to the way in which the training models utilise vectors in waveform prediction, the formatting of the input text has a significant effect on output quality. We found that when using the model as a screen reader for articles from Welsh language news sites, the models far outperformed the shorter sentences that tended to be written by individual testers. A further deterioration can be seen when the language supplied does not conform to standard sentence structures found in either language.

Our experiments with transfer and multi-speaker training to maximise the lesser represented speakers in the dataset gave mixed results with the trained models, verging closer and closer in terms of prosody to that of our largest speaker corpus. Although there are definite improvements that can be made to the training process,

we have demonstrated the potential to train bilingual text-to-speech voices with the BU-TTS corpus and to more efficiently generate new voices with completing only a subset of the full prompt set.

5. Future Work

To further validate this corpus there is potential to train the dataset from phonemes which in many languages produces a higher quality voice. It would also be desirable to complete mean opinion score tests on all of the models generated to ensure a value approaching human speech can be achieved.

6. Conclusion

We presented BU-TTS, the initial instalment of the Bangor University text-to-speech corpus, an open-source Bilingual Welsh-English text-to-speech corpus. Four voices make up the corpus (two female, two male) with roughly 10 hours of recordings. Released under openly permissive CC0 1.0 international license.

7. Acknowledgements

We are grateful to the Welsh Government for funding this work as part of the Text, Speech and Translation Technologies for the Welsh Language project.

8. Bibliographical References

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. (2017). Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR.

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., and Ponti, M. A. (2021). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. *arXiv preprint arXiv:2112.02418*.

Chu, M., Peng, H., Zhao, Y., Niu, Z., and Chang, E. (2003). Microsoft mulan-a bilingual tts system. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Cooper, S., Jones, D. B., and Prys, D. (2019). Crowdsourcing the paldaruo speech corpus of welsh for speech technology. *Information*, 10(8):247.

Gibbon, D., Mertins, I., and Moore, R. K. (2012). *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*, volume 565. Springer Science & Business Media.

Haspelmath, M. and Tadmor, U. (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Ito, K. and Johnson, L. (2017). The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Jones, D. (2020). Macsen: A voice assistant for speakers of a lesser resourced language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 194–201.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Knight, D., Loizides, F., Neale, S., Anthony, L., and Spasić, I. (2021). Developing computational infrastructure for the corcencc corpus: The national corpus of contemporary welsh. *Language Resources and Evaluation*, 55(3):789–816.

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Klimkov, V. (2019). Effect of data reduction on sequence-to-sequence neural tts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7075–7079. IEEE.

Mussakhojayeva, S., Khassanov, Y., and Varol, H. A. (2022). Kazakhtts2: Extending the open-source kazakh tts corpus with more data, speakers, and topics. *arXiv preprint arXiv:2201.05771*.

Myers-Scotton, C. (2017). Code-switching. *The handbook of sociolinguistics*, pages 217–237.

Nilep, C. (2006). "code switching" in sociocultural linguistics. *Colorado research in linguistics*.

Prys, D. and Jones, D. (2018). Gathering data for speech technology in the welsh language: A case study. In *LREC 2018*.

Prys, D. and Jones, D. B. (2019). Wordlists of the most common wordforms in welsh, and the most common english words used in welsh. Jan.

Prys, D., Williams, B., Hicks, B., Jones, D., Ní Chasaide, A., Gobl, C., Carson-Berndsen, J., Cummins, F., Ní Chiosáin, M., McKenna, J., et al. (2004). Wispr: Speech processing resources for welsh and irish. In *Proceedings of the SALTMIL Workshop: First Steps in Language Documentation for Minority Languages*, pages 68–71.

Prys, D., Prys, G., and Jones, D. B. (2016). Cysill ar-lein: A corpus of written contemporary welsh compiled from an on-line spelling and grammar checker. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3261–3264.

Prys, D., Jones, D., Prys, G., Watkins, G., Cooper, S., Roberts, J. C., Butcher, P., Farhat, L., Teahan, W., and Prys, M. (2021). Language and technology in wales: Volume i. 1.

Prys, D., Jones, D., and Prys, G. (2022a). Language and technology in wales: Volume ii. 2. forthcoming.

Prys, D., Watkins, G., and Ghazzali, S. (2022b). Ele d1.34 language report welsh.

Rehm, G. (2014). Digital language extinction as a challenge for the multilingual web. In *Multilingual Web Workshop 2014: New Horizons for the Multilingual Web*. META-NET Madrid.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A comparative handbook*, 55:75.

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Welsh Government. (2018). Welsh language technology action plan.

Zeng, Z., Wang, J., Cheng, N., Xia, T., and Xiao, J. (2020). Alignstts: Efficient feed-forward text-to-speech system without explicit alignment. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6714–6718. IEEE.

Zhiyong, W., Guangqi, C., Meng, M. H., and Cai, L. (2009). A unified framework for multilingual text-to-speech synthesis with ssml specification as interface. *Tsinghua Science and Technology*, 14(5):623–630.