INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Low-Resource Pipeline for Text-to-Speech from Found Data With Application to Scottish Gaelic

Dan Wells,[1] *Korin Richmond,*[1] *William Lamb* [2]

[1]The Centre for Speech Technology Research, [2]Celtic and Scottish Studies
University of Edinburgh, United Kingdom
{dan.wells, korin.richmond, w.lamb}@ed.ac.uk

Abstract

In this work we present an end-to-end pipeline for building a speech corpus and text-to-speech synthesis system for a new language without reference to any expert-defined linguistic resources. We segment and align over 85 hours of Scottish Gaelic recordings found online and select 2- and 8-hour subsets with comprehensive coverage of speech sounds based on self-supervised discrete acoustic unit sequences. We then compare FastPitch models trained on these relatively small data sets using character, acoustic unit and phone inputs. According to native speaker listening test judgements, characters serve well for Gaelic given its regular orthography, even in these limited data scenarios. We release our corpus building recipe so that others may easily apply our work to new languages.

Index Terms: Scottish Gaelic, speech synthesis, low-resource, speech corpus creation, found data

1. Introduction

High-quality neural text-to-speech (TTS) synthesis systems have only been developed for a small proportion of the world's ∼7,000 languages. There are two major issues when building a TTS system for a new language: 1) the availability of suitable speech recordings with matching text transcripts, and 2) the knowledge required to process input text and represent the target language symbolically. The first is a problem of access to data resources, while the second represents the difficulty of constructing linguistic resources for a new target language.

Data requirements for neural TTS have typically been put at some tens of hours of studio-quality speech recordings paired with text transcripts, although recent work has reevaluated these assumptions by switching to non-autoregressive architectures where the burden of learning text-speech alignments alongside acoustic feature prediction is removed [1], or by using powerful self-supervised speech representations to help train some parts of the system on noisier audio data [2]. Much effort has also been directed toward using 'found' data not originally intended for TTS, especially audiobook recordings. Such data generally requires pre-processing, for example segmenting long recordings into individual utterances [3], or filtering to find cleaner subsets for model training [4].

Pronunciation lexicons are one of the key linguistic resources representing a significant investment when building TTS for a new language. Lexicon development requires linguistic knowledge to define phone sets and transcription conventions, and to provide a core set of hand-labelled examples from which to learn a grapheme-to-phone (G2P) model to predict pronunciations for unseen words at synthesis time. Some attempts have been made to avoid this particular dependency by using characters directly as input to TTS systems [3, 5], although the success of this approach may vary depending on how consistent the orthography of the target language is.

Scottish Gaelic is spoken by around 57,000 people in Scotland [6], and is in an interesting position as a minority language with considerable amounts of speech and text data available [7]. This stems from large collections of archival recordings documenting the language, a long history of broadcast media and more recent language revitalisation efforts compiling material to aid learners. There is also an online dictionary providing pronunciations for some 35,000 words [8]. These resources have been used to develop several language technologies for Gaelic, including automatic speech recognition (ASR) [7]. Previous efforts in TTS for Gaelic, however, have either been based on carefully-constructed speech databases in a legacy diphone synthesis context [9], or are proprietary systems (albeit freely available for use in the Scottish public sector) [10].

In this work we develop a replicable and open-source recipe for building a speech corpus and neural TTS system for Scottish Gaelic, all based on publicly-available data.[1] Our work is similar in motivation and method to [3], applied in a more recent TTS framework. We begin by segmenting and aligning over 85 hours of short speech utterances and corresponding text transcripts from long-form audio recordings found online, using a purely character-based acoustic model. We then select utterances up to 2 or 8 hours in total duration to achieve good coverage of speech sounds based on discrete acoustic unit sequences extracted from a HuBERT model pre-trained on English [11]. Finally, we train FastPitch models [12] with these corpora, comparing the performance of character, acoustic unit and phone inputs. With this, we hope to prove an end-to-end process for building a TTS system in a new language with no dependence on linguistic knowledge beyond paired text and speech data.

2. Building a speech corpus without linguistic resources

Our speech data comes from the *Litir do Luchd-ionnsachaidh* 'Letter to Learners' series of Gaelic recordings aimed at intermediate language learners. This series broadcasts on the Gaelic-medium radio station *BBC Radio nan Gàidheal*, with recordings also available on the *LearnGaelic* language-learning website.[2]

Our corpus is based on scraped audio and text transcripts for *Letters* 1–1216. Each recording is around 5 minutes long, so we start with just over 100 hours of unsegmented Gaelic speech. All recordings come from a single speaker, Ruairidh MacIlleathain. Most were recorded at 44.1 kHz and encoded

[1] https://github.com/dan-wells/kiss-aligner/tree/main/egs/learngaelic_litir

[2] https://learngaelic.scot/litir/

10.21437/Interspeech.2023-2056

using Vorbis; for consistency and compatibility with later processing by HuBERT, we reencode all files to 16 kHz, 16-bit mono WAV PCM. The earlier recordings (until *Letter* 307) appear to have used a low-quality microphone and are not suitable for TTS acoustic model training, but all subsequent recordings seem of usable quality. Most recordings also include a short preamble which is not included in the text transcript, and which is sometimes spoken by a second speaker. To prepare for TTS model training, we align and split these long audio files and corresponding transcripts into shorter segments, following an iterative procedure outlined in the following sections.

2.1. Initial segmentation and acoustic model training

We first split the long recordings on silences longer than 1.5 s, to account for the relatively slow and punctuated speaking style used in the *Letters*. We lightly normalise the long text transcripts to encode accented characters consistently and remove some variation in punctuation, then split them into individual sentences. We then calculate the cumulative relative proportions of the respective length of long audio (in seconds) and transcript (in characters) covered by each segment and calculate DTW alignments between the two sequences. We exclude the first audio segment for any recordings which include a preamble. There tend to be more text segments than audio segments per *Letter*, so when processing the DTW alignments we concatenate any text segments aligned to the same audio segment.

We then train an initial Gaelic acoustic model by force-aligning these estimated text-audio pairs. We specify a character-based HMM-GMM alignment model in Kaldi [13], where the 'pronunciation' of each word is given as its constituent characters, avoiding any phone-mapping or language-specific phone set as used in [7]. We expect this to be sufficient for Gaelic given its regular orthography and that our subsequent segmentation process should succeed given only accurate word-level boundaries. We also include punctuation symbols at this stage so that we can maintain them through the following steps. We iteratively train mono- and tied character trigram acoustic models, using a strict beam without retrying failed intermediate alignments, in an attempt to use only data from correctly-estimated text-audio pairs when updating model parameters.

2.2. Full audio alignment and segmentation

Following the Kaldi recipe released by [14], we use our initial acoustic model and a language model trained on the full text transcripts to decode 60 s chunks from the *Letter* recordings. These ASR outputs are aligned against the full reference text using the Smith-Waterman algorithm [15], and the corresponding reference text sections are retrieved based on decoded hypothesis timestamps. This approach can ignore sections in the audio not present in the text, such as the preambles in many of the *Letters*. We restrict discovered segments to be between 5–20 s in duration, with a maximum internal silence length of 1.2 s. We run a final forced-alignment step to verify that the discovered text and audio segments match. This process retrieved and aligned 31,174 segments comprising 86.7 hours of speech.[3]

We maintain the original punctuation in our text transcripts, since it could be useful for downstream TTS models, and because we found it helped to discover sentence-level boundaries during segmentation. Overall, 47.5% of discovered text segments begin with a capital letter and end with clause-final punctuation (*.!?;:*), and 16.4% contain more than one, likely capturing multiple short sentences in a single discovered utterance.

[3]We also tested our alignment pipeline starting from a random selection of 2 or 8 hours of long recordings rather than the full 100 hours, yielding 55 minutes and 5.5 hours of aligned segments, respectively.

Table 1: *Data sets sampled from 86.7 hours of segmented utterances, with triphone coverage relative to the full corpus.*

Data set	# Utts	Avg. duration	Triphone coverage	
2h clean	475	15.1 s	12,376	30.2%
8h clean	2,207	13.0 s	22,995	56.1%
21h noisy	7,335	10.4 s	24,510	59.8%
Validation	380	10.0 s	7,412	18.1%
Test	418	9.7 s	7,112	12.4%

2.3. Acoustically-driven data set selection

With the aim to train TTS models from limited data, we would like to ensure adequate phonetic coverage in small speech corpora. Similar to [4], we approach this data selection problem using purely acoustically-derived features, rather than linguistic features extracted from text using a language-specific frontend, as in [16]. Specifically, we extract discrete acoustic unit sequences from segmented utterances using HuBERT and then select utterances using a greedy algorithm based on unit trigram coverage, replicating [17] as applied to diphone sequences.

We use a HuBERT-BASE model pre-trained on English (namely the 960-hour LibriSpeech train partition [11]) and extract framewise hidden representations from the sixth encoder transformer layer, since intermediate layers have been found to most closely represent phone-level information [11]. We then discretise the continuous hidden representations using a k-means model with 200 clusters learned over all frames extracted from our segmented Gaelic speech data, which we believe should offer enough capacity to capture all relevant phonetic contrasts. While the continuous feature extractor of HuBERT is pre-trained only on English, experiments with similar models have found them to transfer well to unseen languages [18]. Previous work has also found that learning k-means discretisation models on top of HuBERT features is robust to the amount of training data [11], and that these discrete acoustic units correspond closely with phonetic events [19]. These findings give us confidence that this data selection approach should generalise reasonably well across languages, although the tendency to discard pitch information when discretising HuBERT features [20] suggests it may struggle with tonal languages.

Our final set of segmented utterances covers 793 clean *Letters* recordings. We hold out 15 of these each to provide test and validation utterances, and take the top-ranked utterances by acoustic unit trigram coverage from the remainder. We also extract acoustic unit sequences from 21 hours of noisier recordings to supplement a text-to-unit model (see Section 3.3). Table 1 summarises the resulting data sets. Note that our greedy algorithm prioritises longer utterances, giving higher average durations for smaller corpora. We measured triphone coverage in the selected data sets relative to phone sequences generated for the full 86-hour segmented corpus using a Gaelic pronunciation lexicon and G2P model (see Section 3.2). For comparison, we also selected 2h and 8h subsets maximising triphone coverage directly, achieving 37.3% and 68.2% respectively. Relative to these reference sets, our corpora appear to retrieve around 80% of the potential triphone coverage in a given amount of audio, but we might also consider them to reflect contextual variation in speech more directly than corpora selected based on abstracted or perhaps errorful 'reference' phone sequences.

3. Model specification

We train 7 different FastPitch [12] acoustic models using either 2 or 8 hours of speech to predict mel-scaled spectrogram features from character, phone or discrete acoustic unit sequences, as described in the following sections. Following [21, 1], we replace all convolutional layers with depthwise-separable convolutions, reducing overall parameter counts to match our low-data setting. Character- and phone-input models are trained with target durations from forced alignments, while for acoustic unit sequences we derive target durations by run-length encoding repeated consecutive units. We extract target pitch values using the probabilistic YIN algorithm [22]. Each model is trained for 100,000 steps with a batch size of 16. We extract mel-scaled spectrograms from training audio with a frame shift of 320 samples to match the framerate of the HuBERT feature extraction process (50 Hz for 16 kHz audio). We also train a HiFi-GAN (V1) vocoder [23] with matching acoustic feature configuration to generate audio from mel spectrograms. This model is first trained for 270,000 steps (batch size 16) on the English VCTK corpus [24], and then fine-tuned on natural speech from our 8h Gaelic corpus for a further 30,000 steps.

3.1. Character inputs

Gaelic uses the Latin alphabet, excluding the letters ⟨*jkqvwxyz*⟩, with long vowels marked by a grave accent (e.g. short ⟨*a*⟩ vs. long ⟨*à*⟩). Our texts also include some non-Gaelic words (most frequently snippets of English), which we accept as part of using found data and do not attempt to filter out. We preserve any punctuation which might be useful for predicting pauses or utterance-level prosody, plus ⟨'⟩ and ⟨-⟩ which can occur as part of Gaelic words (e.g. the frequently contracted definite article *an* → *a'*). For our FastPitch inputs, we lowercase all characters and mark those at word boundaries with distinct symbols.

3.2. Phone inputs

Am Faclair Beag provides an online Gaelic dictionary, with pronunciations specified for around 35,000 entries [8]. As in [7], we retrieve pronunciations for individual words from this lexicon where possible, and train a word-level 5-gram graphone joint-sequence Sequitur G2P model [25] to handle any out-of-vocabulary items. After removing duplicate pronunciations, we train on 28,000 entries and evaluate on 2,565 held-out words. This model achieves a phone error rate (PER) of 5.7% and word error rate (WER) of 24.1%. We predict the quite closely phonetic Gaelic renderings from *Am Faclair Beag* directly, which represent relatively fine distinctions such as broad vs. slender consonants and vowel nasalisation explicitly. For our model, 71% of incorrect word pronunciations had only a single phone wrong, and mostly within these categories. While these are phonemic distinctions and therefore important to get right, they are also often predictable from context, e.g. with slender consonants appearing between front vowels, and so we might expect our FastPitch model to repair some G2P errors of this kind.

We train a similar model for English on 110,000 words from the Edinburgh-accent surface form of the Unisyn lexicon [26], which we apply only to words containing non-Gaelic characters, since we have no other way to identify non-Gaelic words. Again, we consider this an inevitable source of errors in found data, but note that trying to address it explicitly is both more complicated and less satisfying than a simple character-based approach. As for character inputs, we mark phones at word boundaries with distinct symbols and maintain punctuation.

3.3. Acoustic unit inputs

Our training scheme for discrete acoustic unit inputs is very similar to [27], as we train on ground-truth deduplicated unit sequences extracted from training utterances. At synthesis time, we need to predict unit sequences from text, for which we train a separate text-to-unit (T2U) model similar to [28]. We use an encoder-decoder architecture with two transformer layers each (hidden size 256 throughout, 1024 in transformer feed-forward layers), joined by a general attention module, all as implemented in the *OpenNMT-py* toolkit [29]. We train separate models on utterances from our 8h and 2h data sets, all for 50,000 steps with a batch size of 64.

We train our T2U models on whole utterances rather than isolated words, with the potential benefit of modelling phonological effects across word boundaries directly, without having to devise explicit post-lexical rules as in more traditional G2P frontends. However, this also means that both input and output sequences can be quite long, up to around 300 text characters and 600 deduplicated acoustic units for our data. This makes learning attention alignments difficult, and training with our 2h data set fails completely – the model produces plausibly Gaelic-sounding output with no correspondence to the input text. Training with 8h of data is generally sufficient, although the model sometimes skips words or repeats phrases, especially at the end of utterances which are sentence fragments without final punctuation (an artefact of our segmentation process). In these cases, we have effectively shifted some of the 'babbling'-type problems often observed in attention-based TTS [30] from the acoustic model to the linguistic frontend. Adding 21 hours of text-unit sequences extracted from noisier recordings, which we wouldn't otherwise use for acoustic model training, helps for both 2h and 8h data sets, although it doesn't completely resolve these attention issues. This approach is similar to [2], although based on much less data: up to 28 hours total (similar to per-language T2U training sets in [28]), compared to 3,000 hours. Using a stronger alignment module would likely resolve this problem, for example the monotonic alignment search used in [2] for their *text2vec* model. Apart from these attention issues, we consider this approach to be similar to using character inputs in terms of handling potentially messy found data.

4. Listening test design

Because of the minority status of Gaelic, it is difficult to recruit speakers for subjective evaluation. As such, we designed a listening test based on best-worst scaling (BWS) annotation [31], where we present participants with sets of stimuli synthesised from the same text by 4 different systems and ask them to select the most and least natural-sounding. Each comparison thereby reveals 5 pairwise quality judgements, making it more efficient for gathering responses than AB preference tests which only present a single pair of utterances at a time. It has also been found to yield more consistent results than Likert-type rating scales [31], as commonly used in mean opinion score (MOS) evaluations. Similar to AB tests, we expect the forced-choice nature of this design to enable better discrimination of any significant differences between systems than MOS tests [32].

We were able to recruit 6 Gaelic speakers for our listening test: 3 native speakers and 3 language learners who described themselves as 'confident/intermediate' speakers. Each participant sees 15 BWS questions, each presenting a different test utterance and a different combination of 4 voices. Most test sets contain only TTS voices, but some also include natural speech

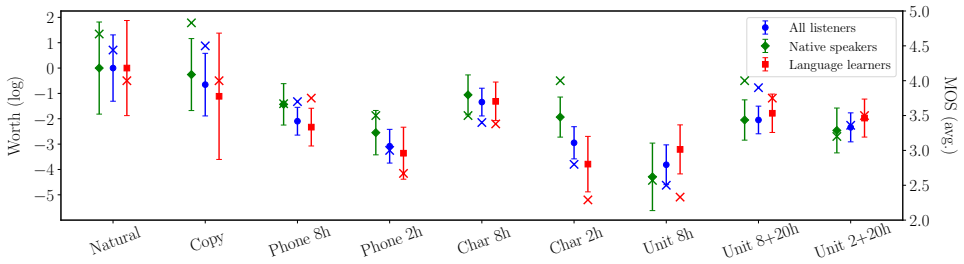

Figure 1: *Subjective ranking of all voices across listener groups from forced-choice comparisons. Each point corresponds to Plackett-Luce model worth parameters, with quasi standard errors where non-overlapping intervals represent statistically significant differences between any pair of voices. For reference, we also show average MOS scores per voice, marked by crosses.*

or copy synthesis samples; since we expect these always to be ranked best where they appear, we include them in a smaller proportion of test tuples in order to focus our data collection on more uncertain judgements between synthetic speech samples. We also include 15 MOS questions to help place the preference rankings in more absolute terms.

We selected 44 sentences from held-out *Letters* for BWS questions (35 TTS-only test tuples, 9 including natural speech and/or copy synthesis) and an additional 90 for MOS questions (10 each for 9 voices, including natural speech and copy synthesis). We selected only full sentences ending in phrase-final punctuation as more natural stimuli to present to participants. This also helped to reduce the incidence of T2U errors for our acoustic unit-input systems. Each BWS sentence was rated by 2 different participants over all listening tests, yielding 45–50 BWS ratings for each synthetic voice and 12 for natural speech and copy synthesis. Each MOS sentence was rated once, giving 10 MOS ratings per voice. Participants took around 20 minutes to complete the listening test, and were paid £5 for their time.

5. Results and discussion

We analyse our BWS preference results using a Plackett-Luce rankings model as implemented in the `PlackettLuce` R package [33]. This implementation accommodates partial subset rankings and ties, matching our BWS setup where each question presents only 4/9 systems, and we have no information about the ranking of the two systems between best and worst. The Plackett-Luce model estimates *worth* parameters for the items of interest, representing the probability that a given item would come first in a ranking of all items. Figure 1 plots each voice according to its log worth relative to natural speech. Also shown are quasi standard errors [34] around this value, where non-overlapping intervals represent significant differences in worth between any pair of voices. We note very few apparent significant differences in these results, which we attribute to the low number of participants. Figure 1 also shows average MOS scores for each system, which are generally consistent with the BWS results, increasing confidence in the apparent system rankings despite the low number of responses.

With 8h of training data, character inputs perform well for both listener groups, and even appear to be slightly preferred over the corresponding 8h phone-input system. We attribute this to the relative transparency of Gaelic orthography, along with possible errors in our G2P frontend. This suggests that character-based TTS could be a good starting point for other languages with regular writing systems, avoiding the effort to compile lexical resources for a new language. As for training corpus size, it seems that 8h is generally preferred over 2h, although the difference may not be as much as expected, especially for character-based systems as judged by native speakers. There is also less of a difference for unit-based systems compared to phone or character inputs, with Unit 2+21h judged almost as good as Unit 8+21h across listener groups. This may be because acoustic unit sequences are a close representation of speech to begin with, so that predicting acoustic features from them is more like a resynthesis task [20], which we could consider a simpler problem than learning an implicit pronunciation model from characters. This in turn suggests that acoustic units could be a good choice for TTS in more limited data scenarios, provided they are paired with a strong T2U model.

There is a notable difference between the 2h character- and 8h unit-input systems when we compare different listener groups, with native speakers and language learners apparently making opposite judgements. On reviewing the samples (one of the authors is a fluent Gaelic speaker), we conclude that the unit-based system tends to have better overall audio quality, while the character-based system tends to match the linguistic aspects of the source text more accurately (perhaps benefitting here from T2U errors). We would expect native listeners to be more sensitive to these linguistic aspects, such as rare word pronunciation, while learners might not pick up on them as readily.

6. Conclusion

In this paper, we presented a full pipeline for developing neural TTS system for Scottish Gaelic, based on publicly-available data and without relying on any expertly-defined linguistic resources. We created a corpus comprising over 85 hours of segmented speech utterances and text transcripts using a character-based acoustic model, then selected smaller data sets with good phonetic coverage according to sequences of self-supervised discrete acoustic units. Subjective listening tests with Gaelic speakers showed that a character-based FastPitch system can achieve good performance given a language with such a consistent orthography, trained on as little as 8 or even 2 hours of found speech data. We also found some indication that using discrete acoustic units as input symbols may be beneficial for TTS in low data scenarios, which we hope to investigate further in future work. We release our corpus building and model code so that others may easily apply our work to new languages.

Acknowledgements: We would like to thank Ruairidh MacIlleathain for allowing us to use his voice for this work. We also thank BBC ALBA and MG ALBA for granting permission as copyright holder and funder of the *Letter to Learners* series of recordings, respectively. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

7. References

[1] A. Pine, D. Wells, N. Brinklow, P. Littell, and K. Richmond, "Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022, pp. 7346–7359.

[2] H. Siuzdak, P. Dura, P. van Rijn, and N. Jacoby, "WavThruVec: Latent speech representation as intermediate features for neural speech synthesis," in *Interspeech 2022*. ISCA, 2022, pp. 833–837.

[3] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: A multilingual corpus of found data for TTS research created with light supervision," in *Interspeech 2013*. ISCA, Aug. 2013, pp. 2331–2335.

[4] P. O. Gallegos, J. Williams, J. Rownicka, and S. King, "An Unsupervised Method to Select a Speaker Subset from Large Multi-Speaker Speech Synthesis Datasets," in *Interspeech 2020*. ISCA, 2020, pp. 1758–1762.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech 2017*. ISCA, 2017, pp. 4006–4010.

[6] National Records of Scotland, "Scotland's Census 2011: Gaelic report (part 1)," 2015. [Online]. Available: https://www.scotlandscensus.gov.uk/media/cqoji4qx/report_part_1.pdf

[7] L. Evans, W. Lamb, M. Sinclair, and B. Alex, "Developing Automatic Speech Recognition for Scottish Gaelic," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*. European Language Resources Association, 2022, pp. 110–120.

[8] M. Bauer, W. Robertson, and E. Dwelly, "Am Faclair Beag." [Online]. Available: https://www.faclair.com/

[9] M. Wolters, "A Diphone-Based Text-to-Speech System for Scottish Gaelic," Master's thesis, Department of Computer Science, University of Bonn, 1997.

[10] CereProc, "CereProc and CALL Scotland Announce World's First Scottish Gaelic Synthetic Voice: Ceitidh," Dec. 2015. [Online]. Available: https://www.cereproc.com/en/CereProc_Gaelic_Synthetic_Voice_Ceitidh

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

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

[14] V. Manohar, D. Povey, and S. Khudanpur, "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 346–352.

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

[16] M. Podsiadło and V. Ungureanu, "Experiments with Training Corpora for Statistical Text-to-speech Systems." in *Interspeech 2018*. ISCA, 2018, pp. 2002–2006.

[17] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.

[18] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised Pretraining Transfers Well Across Languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7414–7418.

[19] D. Wells, H. Tang, and K. Richmond, "Phonetic Analysis of Self-supervised Representations of English Speech," in *Interspeech 2022*. ISCA, 2022, pp. 3583–3587.

[20] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech 2021*, 2021, pp. 3615–3619.

[21] R. Luo, X. Tan, R. Wang, T. Qin, J. Li, S. Zhao, E. Chen, and T.-Y. Liu, "Lightspeech: Lightweight and Fast Text to Speech with Neural Architecture Search," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5699–5703.

[22] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[23] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.

[24] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019.

[25] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[26] S. Fitt, "Unisyn Lexicon." [Online]. Available: https://www.cstr.ed.ac.uk/projects/unisyn/

[27] C. Wang, W.-N. Hsu, Y. Adi, A. Polyak, A. Lee, P.-J. Chen, J. Gu, and J. Pino, "Fairseq S^2: A Scalable and Integrable Speech Synthesis Toolkit," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 143–152.

[28] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless Speech-to-Speech Translation on Real Data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022, pp. 860–872.

[29] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," in *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, 2017, pp. 67–72.

[30] C. Valentini-Botinhao and S. King, "Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech," in *Interspeech 2021*. ISCA, 2021, pp. 2746–2750.

[31] S. Kiritchenko and S. Mohammad, "Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, pp. 465–470.

[32] Y. V. Alvarez and M. Huckvale, "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*. ISCA, 2002, pp. 329–332.

[33] H. L. Turner, J. van Etten, D. Firth, and I. Kosmidis, "Modelling rankings in R: The PlackettLuce package," *Computational Statistics*, vol. 35, no. 3, pp. 1027–1057, 2020.

[34] D. Firth and R. X. De Menezes, "Quasi-variances," *Biometrika*, vol. 91, no. 1, pp. 65–80, 2004.