ABAIR & ÉIST: A demonstration of speech technologies for Irish

Andy Murphy[1]*, Liam Lonergan*[1]*, Mengjie Qian*[2]*, Harald Berthelsen*[1]*, Christoph Wendler*[1]*, Neasa Ní Chiaráin*[1]*, Ailbhe Ní Chasaide*[1]*, Christer Gobl*[1]

[1]Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[2]Department of Engineering, University of Cambridge, UK

murpha61@tcd.ie, llonerga@tcd.ie, mq227@cam.ac.uk, berthelh@tcd.ie, wendlec@tcd.ie, nichiarn@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

The ABAIR initiative [1, 2, 3, 4] has been providing speech technology, and its applications, for Irish language communities for many years. For Irish, as an endangered language, it is particularly important to implement speech technology, not only to offer speakers a resource for language learning and accessibility, but to maintain and preserve the very language itself. The ABAIR initiative has developed linguistic resources, text-to-speech (TTS) systems, automatic speech recognition (ASR), and many applications that utilise these technologies directly. This paper demonstrates and discusses the TTS and ASR technologies created, and provides a glimpse at the applications being developed in parallel.

TTS – ABAIR

The initial aim of the ABAIR (The Irish word for SAY) project was to generate linguistic resources for TTS development. TTS for Irish is a challenge: there is no standard spoken variety, but instead three major dialects, Ulster, Connacht and Munster, and also several sub-dialects. In order for native speakers to be comfortable in the use of TTS, or applications that use it, they must be provided with a system in their own dialect. This has entailed the modular development of linguistic resources, such as letter-to-sound (LTS) rules and text corpora, that could be transposed from one dialect to another. It has also involved the recording of speech corpora of as many speakers as the project's resources have allowed for. Several of these corpora have been recorded in the field, due to the remote location of the native speaking population.

Very early research into Irish synthetic voices used formant and diphone synthesis, but as the technology progressed, so did the voices. The first ABAIR voice released used concatenative unit-selection, and was for the Ulster (Donegal) dialect. It produced natural sounding speech, but concatenation errors yielded occasionally jarring artifacts and its memory requirements made it unsuitable for applications where very fast speech output was required e.g. screen-readers for the visually impaired. This led to the development of statistical parametric-based synthetic voices (HTS in Figure 1) that produce smooth, consistent speech output, with a small memory footprint and computational load, and can yield very fast output. HTS is appropriate, and currently needed, for certain applications, for instance, those based on embedded devices or requiring very low latency. These HTS voices sound less natural than the project's more recent developments (DNN and NEMO voices in Figure 1), which use deep learning-based approaches (the latter uses the FastPitch TTS framework [5]). To make use of the limited amount of data recorded in appropriate conditions, a multi-dialect multi-speaker approach was taken when training the FastPitch voices. All of the recordings for speech synthesis purposes were combined with a corpus of recordings obtained in a number of quite locations (Field recordings for ASR) to create a single corpus approximately 40 hours in duration (see Table 1). The text annotations that accompanied these data were tagged with the appropriate dialect labelling, to ensure that the correct LTS mapping took place. These annotations also included speaker labels, so that the acoustic model could include speaker embeddings. A HiFi-GAN vocoder [6] was trained independently of the acoustic model using the same corpus, and then fine-tuned to each speaker using spectrograms generated from the FastPitch acoustic model.

A synthetic voice preserves a virtual speaker of a dialect, thus preserving the very language itself. To date we have focused on the three main dialects to maximise its general usefulness, but it is also important to create voices for the most endangered dialects. We are currently extending our dialect coverage. Even where relatively few speakers remain in a community, having the local dialect available in the many applications (see below) is a valuable resource for those endeavouring to preserve their dialect. We are also keen to build heritage voices, capturing legendary community figures whose recordings are available. This is the case with the Domhnall Kerry voice, a towering figure who recently died tragically, which we built with his family's full support.

TTS Demo

As shown in Figure 1, users can enter the text that they wish to synthesise in the box provided. They select their desired dialect, speaker, and the method of synthesis. They can also modulate the speech rate and pitch of the synthetic output. The generated files can be saved.

ASR - ÉIST

More recently, Irish ASR, coined as ÉIST (the Irish word for LISTEN), has also been implemented, and made available to the public. A full description of the ASR system development and performance can be found in [7]. The dialect diversity to be catered for compounds the difficulties of the low-resource context of Irish, as the variability must be captured with a limited pool of data. It was a priority from the outset to cater for this cross-dialect variation, and it is an aspiration to build a system that performs equally well for each dialect [8]. The dataset used to train the acoustic model of the ASR system is composed of various speech corpora from different sources (see Table 1) . The speech synthesis corpora served as a starting point for the ASR training. This was supplemented by extensive in the field-recordings. An online MíleGlór recording facility is being used

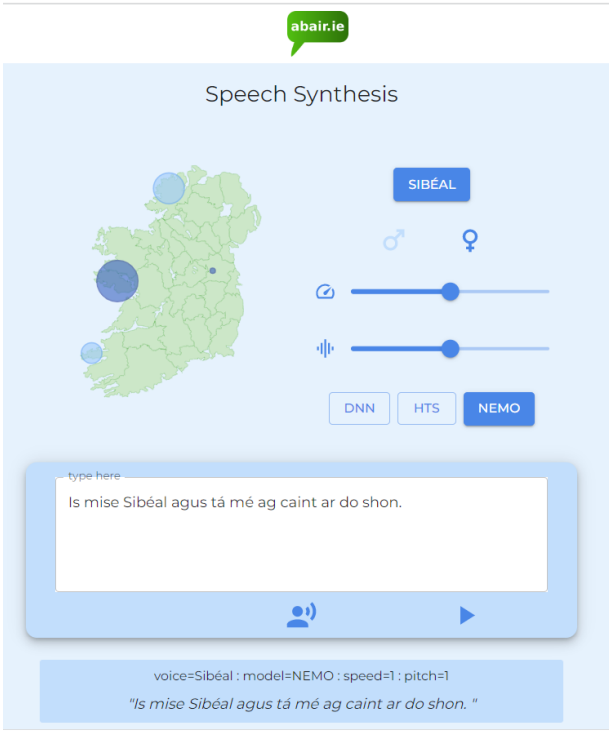

Figure 1: *TTS inteface: Users choose the dialect, gender and synthesis method on the ABAIR webpage (abair.ie).*

for both online and field recordings. External corpora, including broadcast materials, are also being leveraged. Despite the continual growth of our data pool, the speech corpora are still small scale when compared with those commonly used in research or industry. This limits the types of models that can be trained. Initial efforts focused on modular systems, such as hybrid HMM-TDNN models. The letter-to-sound rules developed for synthesis were used to create pronunciation dictionaries and external text corpora were used to train language models for these modular systems [9, 10]. More recently, End-to-End approaches are being investigated, but do not yet compete with our TDNN-HMM system, which with recurrent neural network language model (RNNLM) rescoring achieved WER of 6.4% on a carefully read speech test set, while the best performing End-to-End system, a Conformer model using wav2vec 2.0 as a frontend feature extractor, achieved only a WER of 18%. The RNNLM used was trained with approximately six millions words.

Table 1: *Acoustic datasets used in ABAIR TTS and ASR*

Data type or source	**Duration (hours)**
Speech synthesis recordings	27.3
Audiobooks	32.6
Field recordings for ASR	12.6
Commonvoice	2.3
Broadcast material	143
Total	**217.8**

Current research is investigating recognition systems that can adapt to the dialect or accent of a speaker, by either implicitly or explicitly modelling speaker variety as part of the recognition pipeline. Additionally, our focus is shifting towards leveraging unlabelled speech corpora, which are more readily available than transcribed corpora. Initial experiments in self-supervised pretraining and self-learning have been carried out and will be continued.

ASR Demo

Users can access ÉIST through a web-interface. The user records an utterance, which is then shown as text on the screen. There is also functionality to allow users who wish to save multiple utterances to do so.

Irish Speech Technology Applications

Alongside the development of TTS and ASR is the building of applications for the general public, for education, and for disability and access. Some of these applications can be accessed on the abair.ie website itself, while others are available as downloads from the site. Applications which are currently still under development are also previewed, and are included once the prototypes are sufficiently robust. Public applications include an embeddable web-reader and an android TTS app. Educational apps under development include the An Scéalaí platform [11, 12, 13] for advanced learners, which deploys both TTS and ASR, and Mol an Óige [4], a platform to train phonological awareness and early literacy, intended in the first instance for early learners, but useful to all. Accessibility applications include a downloadable screen-reading facility for the visually impaired [14]. Under development is an AAC communication device [15, 14] for non-speaking users (AAC = augmentative and alternative communication). The public web-reader is also helpful for those with dyslexia.

Future directions

Children's voices will be needed for many educational and accessibility applications, and are a priority for future TTS and ASR research. Bilingual systems will also be required, to enable bilingual users to switch between languages while using the applications.

Acknowledgements

We gratefully acknowledge An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán, which supports the ABAIR and RóbóGlór projects, with funding from the National Lottery as part of Stráitéis 20 Bliain don Ghaeilge, 2010-2030.

References

[1] A. Ní Chasaide, N. Ní Chiaráin, H. Berthelsen, C. Wendler, and A. Murphy, "Speech technology as documentation for endangered language preservation: The case of irish," in *International Congress of Phonetic Sciences*, Glasgow, UK, 2015, pp. 1037–1041.

[2] A. Ní Chasaide, N. Ní Chiaráin, C. Wendler, H. Berthelsen, A. Murphy, and C. Gobl, "The ABAIR Initiative: Bringing Spoken Irish into the Digital Space," in *Proceedings of INTERSPEECH*. Stockholm, Sweden: ISCA, 2017, pp. 2113–2117.

[3] A. Ní Chasaide, N. Ní Chiaráin, H. Berthelsen, C. Wendler, A. Murphy, E. Barnes, and C. Gobl, "Can we defuse the digital timebomb? linguistics, speech technology and the irish language community," in *Proceedings of the Language Technologies for All (LT4All)*, Paris, France, 2019, pp. 177–181.

[4] A. Ní Chasaide, N. Ní Chiaráin, H. Berthelsen, C. Wendler, A. Murphy, E. Barnes, and C. Gobl, "Leveraging Phonetic and Speech Research for Irish Language Revitalisation and Maintenance," in *ICPhS 2019: the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019, pp. 994–998.

[5] A. Lancucki, "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.

[6] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[7] L. Lonergan, M. Qian, H. Berthelsen, A. Murphy, C. Wendler, N. Ní Chiaráin, C. Gobl, and A. Ní Chasaide, "Automatic speech recognition for Irish: the ABAIR-ÉIST system," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, 2022, pp. 47–51. [Online]. Available: https://aclanthology.org/2022.cltw-1.7

[8] L. Lonergan, M. Qian, N. Ní Chiaráin, A. Ní Chasaide, and C. Gobl, "Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer?" in *Proceedings of INTERSPEECH*, Dublin, Ireland, in press.

[9] L. Lonergan, M. Qian, N. Ní Chiaráin, C. Gobl, and A. Ní Chasaide, "Cross-dialect lexicon optimisation for an endangered language asr system: the case of irish," in *proceedings of INTERSPEECH*, 2022. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2022-838

[10] M. Qian, H. Berthelsen, L. Lonergan, A. Murphy, C. O'Neill, N. Ní Chiaráin, C. Gobl, and A. Ní Chasaide, "Automatic speech recognition for irish: testing lexicons and language models," in *Proceedings of the 33rd Irish Signals and Systems Conference (ISSC)*, Cork, Ireland, 2022.

[11] N. Ní Chiaráin and A. Ní Chasaide, "An scéalaí: Synthetic voices for autonomous learning." in *P. Taalas, J. Jalkanen, L. Bradley and S. Thouësny (eds.), Future-proof CALL: Language Learning as Exploration and Encounters - short papers from EUROCALL 2018*. Jyväskylä, Finland: Research-publishing.net, 2018, pp. 230–235.

[12] ——, "The potential of text-to-speech synthesis in computer-assisted language learning," in *Alberto Andujar (ed.), Recent Tools for Computer and Mobile-Assisted Foreign Language Learning*. Hershey, PA: IGI Global, 2020, pp. 149–169.

[13] N. Ní Chiaráin, M. Comtois, O. Nolan, N. Robinson-Gunning, J. Sloan, H. Berthelsen, and A. Ní Chasaide, "Celtic CALL: Strengthening the Vital Role of Education for Language Transmission," in *Proceedings of the 4th Celtic Language Technology Workshop, CLTW 2022 at Language Resources and Evaluation Conference, LREC 2022*, Marseille, France, 2022, pp. 71–76.

[14] A. Ní Chasaide, E. Barnes, N. Ní Chiaráin, R. McGuirk, O. Morrin, M. Nic Corcráin, and J. Cummins, "Challenges in assistive technology development for an endangered language: an Irish (Gaelic) perspective," in *SLPAT 2022 - 9th Workshop on Speech and Language Processing for Assistive Technologies, Proceedings of the Workshop*, 2022, pp. 80–87.

[15] E. Barnes, O. Morrin, A. Ní Chasaide, J. Cummins, H. Berthelsen, A. Murphy, M. Nic Corcráin, C. O'Neill, C. Gobl, and N. Ní Chiaráin, "AAC don ghaeilge: the prototype development of speech-generating assistive technology for Irish," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 127–132.