INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

All Together Now: The Living Audio Dataset

David A. Braude[1], *Matthew P. Aylett*[1], *Caoimhín Laoide-Kemp*[1], *Simone Ashby*[2], *Kristen M. Scott*[2], *Brian Ó Raghallaigh*[3], *Anna Braudo*[4], *Alex Brouwer*[4], *Adriana Stan*[5]

[1]CereProc Ltd., UK
[2]University of Madeira, Madeira ITI, Portugal
[3]Dublin City University, Ireland
[4]University of Edinburgh, UK
[5]Technical University of Cluj-Napoca, Romania

{dave,matthewa}@cereproc.com, laoidekc@tcd.ie, {simone.ashby,kristen.scott}@m-iti.org, brian.oraghallaigh@dcu.ie, {annabraudo,c.alexbrouwer}@gmail.com, adriana.stan@com.utcluj.ro

Abstract

The ongoing focus in speech technology research on machine learning based approaches leaves the community hungry for data. However, datasets tend to be recorded once and then released, sometimes behind registration requirements or paywalls. In this paper we describe our Living Audio Dataset. The aim is to provide audio data that is in the public domain, multilingual, and expandable by communities. We discuss the role of linguistic resources, given the success of systems such as Tacotron which use direct text-to-speech mappings, and consider how data provenance could be built into such resources. So far the data has been collected for TTS purposes, however, it is also suitable for ASR. At the time of publication audio resources already exist for Dutch, R.P. English, Irish, and Russian
Index Terms: dataset, audio, multilingual, crowd building

1. Introduction

In the modern era of TTS research the focus has primarily been on machine learning approaches, either using parametric systems such as Merlin [1] and Idlak [2], or directly predicting waveforms as in Wavenet [3]. In the context of TTS, machine learning relies on 'ground truth' data to be trained. Even older machine learning techniques such as HMMs in HTS [4] or MaryTTS [5] need sufficient data to create models. This data requirement is no less true for speech recognition.

In this dataset we wish to involve non-technical communities in building the resources, particularly those communities where minority languages or non-standard accents are spoken. For example while there is an abundance of audio data available for English, none is available for the variety of English spoken on Bere Island (Ireland). Developing language resources for an island with a population of less than 200 people is not practical in most circumstances. However, if the right tools are available for the community to record the data themselves this opens up avenues for regional data to be collected. By providing not only the data but also a corresponding set of tools we aim to *crowd build* the dataset with communities which stand to gain from having access to more language resources and the technological outputs that follow from such efforts. Note that this is in contrast to *crowd sourcing*, which implies top-down data gathering, such as might be performed by a research team evaluating a bespoke spoken language technology application.

We not only want to include communities in building a repository of spoken language data, we also aim to improve existing language resources, such as pronunciation lexicons, normalisation rules, and the provision of text corpora. While it is not yet clear how to build the tools for non-experts, by providing a basic format we can start the data gathering process. By establishing a flexible format, many different methods can be trialled for collecting language resources, whose data may be utilised regardless of whether the crowd building effort gains traction.

Recently, powerful machine learning approaches to speech synthesis and speech recognition have called into question the value of linguistic resources such as pronunciation lexicons. In such cases, NLP practitioners often opt for a graphemes-only approach (or even Unicode byte strings) to train systems and implicitly derive underlying phonemic representations, with the logic that *"such a system alleviates the need for laborious feature engineering, which may involve heuristics and brittle design choices"* [6]. Other state-of-the-art systems, such as Tacotron, opt for text-audio pairs as input [6]. While producing good results, these systems require vast amounts of training data for deriving phonemic representations, and fixing incorrectly generated pronunciations can be difficult.

Without diminishing the utility of these systems, we argue that low resource languages are best addressed through a combination of machine learning and crowd building approaches. Moreover, linguistic resources such as pronunciation lexicons offer a relatively straightforward means of updating, correcting and adding pronunciations, and dealing with loan words, irregular pronunciation, invented words, and named entities. Within this context, letter-to-sound (LTS) conversion offers a robust means of handling out-of-vocabulary (OOV) words, while simple hands-on procedures for correcting or adding lexical entries are regarded as an explicit design feature.

The Living Audio Dataset (LADs) is by no means the first of its kind, however many alternative resources have significant barriers [7]. Some, such as TIMIT [8], are not suitable for TTS as there are not enough recordings of each speaker. Others, such as the data used in Blizzard challenges [9], have not been released into the public domain, or lack uniform file formats across languages [10]. Even recent crowd sourcing efforts continue to remain under the control of the researcher, instead of allowing the corpus to transform and grow through a process of commons-based peer production [11]. The predominate existing datasets used by TTS researchers - LJ speech [12] and CMU Arctic [13] - are static.

In contrast, LADs offers a platform for continuous expan-

http://dx.doi.org/10.21437/Interspeech.2019-2448

sion of the available audio data that can be used for any purpose. Additionally, unlike most corpora, LADs has publicly available and maintainable documentation focusing on its exploitation [7], enabling individuals and communities to expand the dataset without necessarily requiring the actions of a researcher to initiate data gathering.

Currently the dataset contains lexicons, recording scripts and audio data for Dutch, R.P. accented English, Irish, and Russian. These resources have been curated and collected in professional recording studios by both language experts and Masters level students. The language resources and audio are both in the public domain. The language resources which were created for use with the Idlak TTS system can be found in the Idlak GitHub repository. The audio data and recording scripts are maintained in the LADs repository.

Treating language resources (audio, lexicons, text corpora) as dynamic rather than static requires a change of practice. Currently it is difficult to even find corpora, though efforts have been made to try and address this [7, 14], while typically only annotations and tools are regards as changeable [15]. This presents us with a challenge. Fortunately, researchers and engineers working in so-called *big data* have already begun addressing issues such as data *heterogeneity* and *inconsistency and incompleteness* (see [16] for an introduction). A key concept in building dynamic languages resources is the provenance or origin of the new or edited information. The ability to rapidly generate a new pronunciation lexicon using a tried and tested letter-to-sound (LTS) system offers an important means of bootstrapping languages. When that lexicon is shared and crowd building is used to improve entries, its provenance becomes important for establishing trust in the information. For example one is likely to have more faith in pronunciations that have been hand-annotated by an L1 expert over those generated by an LTS system.

In the sections that follow we describe the types of resources are currently available as part of LADs and how they were gathered. We then discuss how the dataset can be augmented, expanding the existing set of tools for creating new languages and reaching out to new speakers. Finally, we describe how the dataset and tools can be accessed.

2. Current resources

2.1. Language resources

The first stage in creating language resources for systems that aim to convert text into speech or speech into text is to collect some text. This is more problematic than it might seem due to copyright barriers and the fact that public domain texts are often archaic or technical. A good starting text corpus provides a critical resource for determining word frequency, likely normalisation issues, examples of homographs and abbreviations, and so on. By taking the top most frequent 30k words, one can establish a reasonable basis for bootstrapping a pronunciation lexicon. For TTS it is also the best means for generating a recording script to collect audio data.

Pronunciation lexicons are accent-specific. A variety of approaches have been used to build lexicons. Generally, anything goes for bootstrapping a lexicon, from applying open-source LTS rules, to seed lexicons composed of high-frequency entries together with machine learning LTS approaches, to the use of open resources such as Wiktionary[1]. Our lexicons use the phonesets for Idlak Tangle [2] which are based on the those developed by CereProc Ltd. for their commercial TTS system.

[1] https://www.wiktionary.org/

Table 1: *Languages and accents.*

Language	Accent	Lexicon	Script
Irish (ga)	Gen. Irish (ie)	Yes	Yes
Dutch (nl)	Netherlands (nl)	Yes	Yes
English (en)	Gen. American (ga)	Yes	Yes
	Received Pron. (rp)	Yes	
Romanian (ro)	Romania (ro)	Yes	Yes
Russian (ru)	Russia (ru)	Yes	Yes

For this dataset we have created phonetically balanced recording scripts using modern language sources. The Dutch, English, and Russian scripts were generated from Wikipedia articles. The Irish script was generated from the *Corpas na Gaeilge Comhaimseartha* (Corpus of Contemporary Irish) [17]. The Romanian script was developed using the The SWARA Speech Corpus [18]. Not all accents have unique recording scripts. However, in most cases using a recording script from a different accent will still produce audio with good phonetic coverage. Table 1 shows the languages and accents currently maintained in LADs and which resources are available for each.

2.2. Audio

The audio data breakdown at the time of writing is given in Table 2. Speaker-specific versions of the recording script in SSML are available for overriding pronunciations if needed. As this is a living dataset, additional audio and languages may be added. The README accompanying the dataset will be kept up-to-date with this information. With the exception of the RP English speaker who was a paid actor, the data recorded so far has been collected from volunteers in Edinburgh without a casting processes [19].

3. Collecting new data

The current section describes how to create new resources within LADs. It is divided into language resources and speakers. The process of starting a new language does require more expertise than recording new speakers, however, we have developed more tools for assisting in language resource creation (see Section 3.3). The entire process is summarised in Figure 1.

3.1. New languages and accents

LADs uses ISO–639–1 two-letter language codes.There is no standardised list of accent codes so it is left to contributors to adopt their own. Like the language codes, the accent codes are also composed of two letters, which may overlap with language codes provided they are unique within a language.

The first step in developing for a new language is creating or finding a corpus of text. If you are creating a new corpus and copyright allows, there is a format for uploading text sources to the dataset. A text corpus allows you to build a word frequency list. The word frequency list then forms the basis of the lexicon. As indicated one possible source of text data is Wikipedia, for which there are tools for downloading text-based content into corpora. Wikipedia should be used with caution, as some languages may only have very short articles or articles not written by fluent users of the language.

A speech synthesis phoneset has different requirements from that used in traditional linguistic descriptions. Often allophonic variation can be handled by the synthesis system with-

Table 2: *Available audio data, shown as it appears in the Living Audio Dataset README.*

Speaker	Language	Accent	Gender	Total duration(mm:ss)	Sample rate (Hz)
ABW	Dutch (nl)	Netherlands (nl)	Male	57:49	48 000
RBU	English (en)	Received Pron. (rp)	Male	50:50	48 000
CLL	Irish (ga)	Non-native (ie)	Male	61:56	48 000
ABR	Russian (ru)	Russian (ru)	Female	34:58	48 000

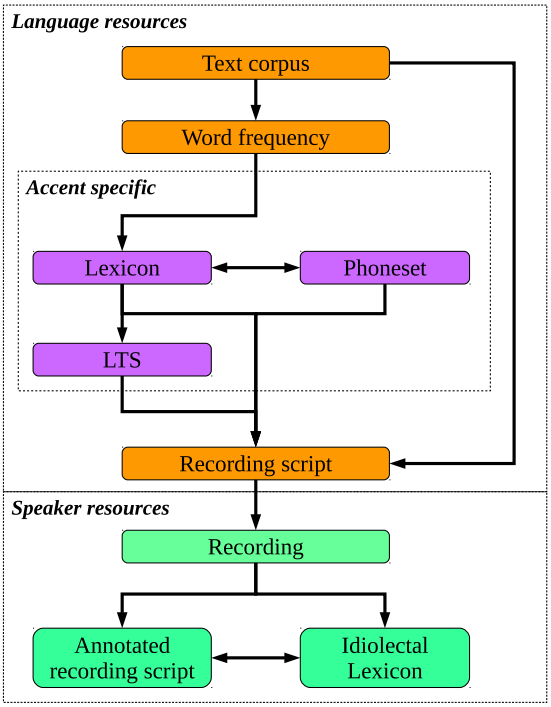

Figure 1: *Workflow for building new resources. Language resources may have accent-specific versions.*

out specific phones to represent these alternate sounds. Legacy and commercial TTS systems may be happy to share a phoneset for a language, and it is better to use one that has proven appropriate in the past. Once the lexicon takes shape LTS rules or models can be trained. For existing lexicons we have used Phonetisaurus [20] for the LTS training. Our experience is that at least 1,000 words are necessary to create an initial LTS model using Phonetisaurus. The lexicon can be further bootstrapped using a combination of Wiktionary and the tools described in Section 3.3. A Wiktionary generated lexicon will have errors and need to be corrected by a speaker of the language.

Using the lexicon, LTS rules and text corpus, the recording script can be generated. There are a few issues with recording scripts that should be considered: firstly they should have complete phonetic coverage, preferably in many contexts (e.g. word / phrase initial and final positions); secondly the utterances should be easy to read and not too long; and thirdly they should be in the language of the recording script. We have provided a tool (described in Section 3.3) that can parse LADs text corpus format and generate a new recording script. However, it still needs to be manually edited to ensure it meets the requirements above. For example, while LTS rules will perform reasonably well at handling unusual phoneme combinations and outputting the names of the Polish football team players, it will not be easy for a non-Polish speaker to correctly pronounce them.

One requirement of the recording script is that it must be fully normalised. In other words all numbers must be written out as words, and acronyms must be expanded into separate uppercase letters. It is also encouraged that all words in the recording script be added into the lexicon, even if the LTS rules correctly predict them. It is very important to perform a manual check of these entries to avoid variability in how these acronyms are expanded. An example from the General American English recording clearly shows this normalisation:

```
<fileid id="z0001_015">
        To eight thirty P M.
</fileid>
```

The file identifiers are given in the recording script with the `fileid` property. This is how audio files will be identified. The `fileids` follow a fixed pattern: first is a single lowercase ASCII letter, denoting the genre (speaking style, purpose or source type). For instance, `z` is used for phonetic coverage, `q` represents questions, and `n` indicates that the corresponding content has been sourced from news feeds. This is followed by a four-digit, zero-padded number, used to create groups within a genre. Examples include a paragraph index, or numeric references to the individual news articles. Lastly, a three-digit, zero-padded number is used for uniquely identifying each utterance. For example, `z0001_001` is the first utterance in the phonetic coverage genre. Other than the use of `z` for phonetic coverage, no other genres are predefined.

3.2. New speakers

Collecting audio from a new speaker using an existing recording script is, by design, a simple process. The requirements are that separate audio files must be created for each utterance in the recording script with file names matched to the `fileid` property. It is not required that the entire recording script be used. Instead, it is suggest that users start at the top and work down to maximise phonetic coverage (if the tool in Section 3.3 was used). MaryTTS [5] provides RedStart which can be used for recording.

Ideally audio is captured in a recording studio or anechoic chamber, however, a quiet room with soft furniture to reduce reverberation could suffice. It is also important to record at the highest possible quality allowed by the microphone. Finally, the recording software and microphone should have any compression disabled.

Each speaker is assigned a three letter informant code. Each audio file should have the informant code, followed by an underscore and the `fileid`. For example, if the speaker is `abw` file names should be of the form `abw_z0001_001.wav`.

Once the recording is complete the audio is simply uploaded to the Internet Archive[2] in a zip file and the url and speaker information are added to the git repository. A copy of the script is placed with the speaker in case the script changes, or there are fixes to the pronunciation, as described below.

Pronunciation issues can be fixed in one of two ways. If a single word has been mispronounced, it can be corrected in the speaker-specific script using the `lex` XML tag. For exam-

[2] `https://archive.org/`

ple if in the US English phrase *But there are some bright spots* the speaker had dropped the final *t* sound in 'bright' it could be fixed with

```
<lex phon="b r ay1">bright</lex>
```

Note that such solutions may not be practically applied for more idiolectal pronunciation variations, in which cases one should create a copy of the lexicon for the speaker and replace the word there. In either case the new pronunciation(s) should subsequently be uploaded to the repository.

If the language resources are available in Idlak, and the README file has been updated in LADs, then a Idlak-Tangle TTS model can be trained using by using the existing LADs example. All that should be needed is specifying the new speaker code along with the language and accent. For a single speaker at least 400 utterances should be recorded (depending on the phonetic complexity of the language) for training a Tangle voice.

3.3. Tools

To assist in the collection of new data we have created some tools that have been released alongside the dataset. These tools are currently focused on generating resources for new languages rather than collecting audio for, and improving existing languages, which will be the focus of some future work. All the tools are written in Python 3.

The first tool parses Wiktionary for the given language and generates a lexicon. As the format is not uniform across languages the tool has been designed to be easily extended to new languages. The tool outputs a lexicon in the Idlak format where the pronunciations are written using the International Phonetic Alphabet (IPA).

The second tool converts lexicon entries that have IPA pronunciations to Idlak phonesets. This is done with a specified mapping. The tool converts the longest possible IPA strings to the phones from the phoneset. An example mapping is:

```
<map pron="ax" ipa="aʊ"/>
```

Which would take precedence over

```
<map pron="a" ipa="a"/>
```

because the IPA string is longer. It should be noted that multiple phonemes can be in the `pron`. To help with the mapping, a companion tool may be used for creating a default mapping from the Idlak phonesets.

The final tool uses the pronunciation lexicon and selected text corpora to create a recording script. Pronunciations are generated for every word in the text corpora, either using the lexicon or through LTS models trained on the lexicon. The LTS system is part of the Idlak front-end and must be compiled separately. The generated recording script maximises phone-to-phone co-articulation coverage. To assist in developing text corpora there is a small utility for downloading articles from Wikipedia in the desired language, however any text source in the appropriate format could be used.

4. Future work

At this point we have assembled all of the necessary tools to enable experts to create new languages, accents, and record new speakers. The next step is to create a set of user-friendly interfaces for engaging community members in contributing to the dataset generation effort, which largely includes recording new audio and improving lexicon entries.

One of the most difficult parts of collecting data of this type is ensuring that volunteers understand the risks of having their audio uploaded into the public domain. We will provide a suggested information sheet and informed consent process to help facilitate a communication process that complies with ethical standards [21].

A key challenge for the Living Audio Dataset is to build in a system for provenance. If a word is corrected how confident can we be that the correction is valid? If extra words are added to Wiktionary, will these be better than pronunciations we may have generated by algorithm or using legacy resources? If we can determine provenance we can also use this information to make crowd building more effective, e.g. asking the community to improve the pronunciation of low-confidence entries. Ideally trusted crowd-built resources in LADs may then be automatically fed back into other resources such as Wiktionary.

5. Obtaining the dataset

The audio, recording scripts, and corpus tools are available from GitHub via the following link:

`https://github.com/Idlak/Living-Audio-Dataset`

It should be noted that the actual audio data is hosted on the Internet Archive[2]. The GitHub repository contains direct links to zipped versions of the audio at their original recording rate and not the audio itself, so cloning the repository does not require much space.

As the pronunciation lexicons are intended for use with Idlak Tangle and are written using the Idlak phonesets, these are hosted on Idlak:

`https://github.com/Idlak/idlak`

Unless otherwise noted in the repository, everything is considered in the the public domain.

6. Conclusion

The Living Audio Dataset not only provides new multilingual data for the speech research community, but also a platform for expansion. By following the processes highlighted in this paper the data follows a uniform format, making it easier to use these resources yo create new TTS voices. One has simply to add the resources to the dataset to be able to train a new voice with Idlak Tangle.

Language is a key facet in the identity of a geographic community. Having access to Information and Communications Technologies (ICTs) that enable communication in one's native language or dialect is a fundamental step in addressing the technology gap that currently divides dominant and minority language speakers. With the possibility for rapid results minority language speakers can enjoy the the sense of ownership that comes with crowd building commons-based assets that can help level the playing field in terms of access to ICTs, and thus opportunities. We hope that this, in turn, will encourage more people to donate their voices to help keep their language or dialect alive.

Implicit in the design of this resource and platform is the fact that the data will gradually change over time to reflect changes in contemporary use. Language is not static and nor should our resources be.

7. Acknowledgements

This work was supported by the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No 780890 (Grassroot Wavelengths).

This work was partially supported by a grant of the Romanian Ministry of Research and Innovation, project number PN-III-P1-1.2-PCCDI-2017-0818/73.

8. References

[1] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System." in *SSW*, 2016, pp. 202–207.

[2] B. Potard, M. P. Aylett, D. A. Braude, and P. Motlicek, "Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN." in *Interspeech*, 2016.

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[5] S. Le Maguer and I. Steiner, "The "Uprooted" MaryTTS entry for the Blizzard Challenge 2017," in *Blizzard Challenge*, Stockholm, Sweden, Aug. 2017. [Online]. Available: http://festvox.org/blizzard/bc2017/MaryTTS_Blizzard2017.pdf

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Interspeech*, 2017.

[7] N. Calzolari, C. Soria, R. Del Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani, and S. Piperidis, "The LREC map of language resources and technologies," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[8] J. S. Garofolo, L. Lamel, W. M. Fisher, J. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 11 1992.

[9] S. King, "Submissions and listening test results from previous Blizzard Challenges," http://www.cstr.ed.ac.uk/projects/blizzard/data.html, accessed: 22 Mar 2019.

[10] L. Ha, "Crowdsourcing a Text-to-Speech Voice for Low-Resource Languages," https://ai.googleblog.com/2015/09/crowdsourcing-text-to-speech-voice-for.html, accessed: 4 April 2019.

[11] J. Smith, A. Tsiartas, V. Wagner, E. Shriberg, and N. Bassiou, "Crowdsourcing emotional speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5139–5143.

[12] K. Ito, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[13] J. Kominek and A. W Black, "The CMU Arctic speech databases," *SSW5-2004*, 01 2004.

[14] N. Ide, J. Pustejovsky, N. Calzolari, and C. Soria, "The SILT and FlaReNet International Collaboration for Interoperability," in *Proceedings of the Third Linguistic Annotation Workshop*, ser. ACL-IJCNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 178–181. [Online]. Available: http://dl.acm.org/citation.cfm?id=1698381.1698415

[15] A. Witt, U. Heid, F. Sasaki, and G. Sérasset, "Multilingual language resources and interoperability," *Language Resources and Evaluation*, vol. 43, no. 1, pp. 1–14, Mar 2009. [Online]. Available: https://doi.org/10.1007/s10579-009-9088-x

[16] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86–94, 2014.

[17] K. Ní Loingsigh, B. Ó Raghallaigh, and G. Ó Cleircín, "The Design and Development of Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish)," in *9th International Corpus Linguistics Conference (CL2017)*, 2017, pp. 113–117.

[18] A. Stan, F. Dinescu, C. Tiple, S. Meza, B. Orza, M. Chirila, and M. Giurgiu, "The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset," in *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, July, 6-9 2017.

[19] A. Braudo, "Russian language voice building for Idlak TTS System," M.Sc. dissertation, 08 2018.

[20] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[21] K. Scott, S. Ashby, D. A. Braude, and M. P. Aylett, "Who owns your voice? Ethically sourced voices for non-commercial TTS applications," in *Conversational User Interfaces (CUI2019)*, 2019, accepted.