Now You See Me, Now You Don't: 'Poverty of the Stimulus' Problems and Arbitrary Correspondences in End-to-End Speech Models

Daan van Esch
Google Speech
1600 Amphitheatre Parkway, Mountain View, CA 94043
dvanesch@google.com

Abstract

End-to-end models for speech recognition and speech synthesis have many benefits, but we argue they also face a unique set of challenges not encountered in conventional multi-stage hybrid systems, which relied on the explicit injection of linguistic knowledge through resources such as phonemic dictionaries and verbalization grammars. These challenges include handling words with unusual grapheme-to-phoneme correspondences, converting between written forms like '12' and spoken forms such as 'twelve', and contextual disambiguation of homophones or homographs. We describe the mitigation strategies that have been used for these problems in end-to-end systems, either implicitly or explicitly, and call out that the most commonly used mitigation techniques are likely incompatible with newly emerging approaches that use minimal amounts of supervised audio training data. We review best-of-both-world approaches that allow the use of end-to-end models combined with traditional linguistic resources, which we show are increasingly straightforward to create at scale, and close with an optimistic outlook for bringing speech technologies to many more languages by combining these strands of research.

Keywords: speech recognition, speech sythesis, end-to-end modeling, text normalization, pronunciation modeling

1. Introduction

In recent years, so-called 'end-to-end' models have become increasingly popular for both automatic speech recognition (ASR) and text-to-speech (TTS) applications. The precise meaning of 'end-to-end' is not exactly defined, and there are many variations on the theme, but in the case of ASR, end-to-end models are typically understood to be all-neural systems that take in audio features and emit some subword-level unit like bytes (Li et al., 2019), graphemes, or wordpieces directly (Prabhavalkar et al., 2017). Such all-neural systems side-step the need for a manually-curated phonemic dictionary of the target language (Sainath et al., 2018; Kim et al., 2020), and contrast with 'conventional' or 'hybrid' approaches, which consist at least partially of non-neural components, like a phonemic lexicon finite-state transducer (Mohri et al., 2002) or hand-written verbalization grammars (Sak et al., 2013) to turn words like 'twelve' into '12'. Similarly, in TTS, end-to-end architectures like Tacotron (Wang et al., 2017) can take in graphemes and emit audio, again without going through intermediate phases that are common in conventional TTS systems, such as text normalization to turn input like '12' into 'twelve' (Sproat et al., 2001), and grapheme-to-phoneme conversion, typically employing a combination of phonemic dictionaries and machine-learning models (Bisani and Ney, 2008).

In conventional speech processing systems, with multiple individual components involved, jointly optimizing over the entire system from start to finish is hard, and errors may compound, which can be detrimental to quality (Wang et al., 2017). By design, all-neural end-to-end approaches do not have this issue, while also being much simpler to train (Chiu et al., 2017), as well as being significantly smaller in terms of disk size, enabling deployment of high-quality models on devices like smartphones (Kim et al., 2020). In addition, being able to avoid the need for injecting linguistic knowledge, such as phonemic dictionaries or verbalization grammars, is frequently considered to be an advantage of end-to-end modeling approaches—for example, (Wang et al., 2017) points out that such components require 'extensive domain expertise and are laborious to design', while an end-to-end approach 'alleviates the need for laborious feature engineering'. Similarly, (Sotelo et al., 2017) argues that end-to-end modeling for TTS 'eliminates the need for expert linguistic knowledge, which removes a major bottleneck in creating synthesizers for new languages'.

Unarguably, developing such linguistic resources can take a non-negligible amount of effort, even despite much progress on tools and methodologies that can help alleviate this burden (Kim and Snyder, 2013; Rutherford et al., 2014; Deri and Knight, 2016; Lee et al., 2020; Ritchie et al., 2019; Bleyan et al., 2019; Ritchie et al., 2020). And it is, of course, desirable to mitigate bottlenecks in developing speech systems in whatever way possi-

Second Workshop on Computation and Written Language (CAWL 2024) @LREC-COLING-2024, pages 43–52

ble: if it were indeed possible to completely avoid the need for linguistic resources while still building systems that are in every way just as capable of handling ASR or TTS tasks as conventional systems, or even better at doing so (as suggested by Sainath et al. (2018)), that would be excellent. However, this may be infeasible, as the relationship between the spoken and the written form of human languages is typically far from straightforward: such correspondences frequently turn out to be full of entirely arbitrary phenomena that cannot be derived, or that would be very difficult to derive, through generalization from a randomly selected large set of training data in the target language.

We provide an overview of such challenging correspondences, and argue that this area has not been receiving enough attention and analysis in the literature on end-to-end systems, calling for more research and analysis along the lines of (Fong et al., 2019; Taylor and Richmond, 2019) to investigate how well end-to-end speech modeling approaches are equipped to deal with such problems. We focus on a practical description of the problem space and common mitigation techniques more so than empirical experiments, since the ability of end-to-end speech systems to handle such arbitrary correspondences will differ depending on factors such the composition of the training data, plus model architecture and size. We argue that these issues will become all the more relevant as end-to-end modeling approaches are adopted that use only minimal amounts of supervised target-language training data: for example, (Baevski et al., 2020; He et al., 2021) report impressive progress on extending speech technologies to new languages using just minutes of transcribed target-language audio, but the resulting models are likely to struggle with such arbitrary correspondences when no linguistic resources are used.

2. Poverty of the Stimulus

As an example, assume for the sake of argument that an end-to-end model's training data contains *all* numbers in the English language except for '12' (and except numbers of which '12' is a constituent, like '112' as 'one hundred and twelve'). Given this data, this hypothetical model would be unable to know that 'twelve' should be transcribed as '12' in ASR tasks, or that '12' can be read as 'twelve' in TTS tasks. This is because there is no possible generalization that would let the model determine that the written form '12' corresponds to the spoken form 'twelve'. To avoid generalizing to something incorrect like 'twoteen' (by analogy with 'fourteen', 'sixteen', 'seventeen', and so on) or 'ten-two' (by analogy with 'twenty-two', 'thirty-two', and so on), an end-to-end model *needs* to observe at least one example of this arbitrary correspondence in its training data.

This problem is by no means limited to '12' alone, of course: even just among English numbers between 10 and 20, cases like '11' and 'eleven', '13' and 'thirteen', and '15' and 'fifteen' are not quite straightforwardly generalizable either. More generally, such examples are somewhat reminiscent of what's known in linguistics as the 'poverty of the stimulus', which is the argument that children are not exposed to enough information to correctly induce the rules of their native language, and that it must therefore be the case that the human brain must contain some form of innate knowledge about how languages work (Laurence and Margolis, 2001). This debate is far from settled in linguistics, but what is relevant to us here is the idea of analyzing a learner's input to understand the limits of generalization based on such training data, and to test how well learners generalize at various points of the learning process.

We believe end-to-end speech processing would benefit from (1) taking into account the limits of generalization when it comes to correspondences between spoken and written forms of language, (2) understanding to what extent the training data used for a given model can theoretically allow it to learn these correspondences, (3) evaluating the degree to which this process is successful in practice, e.g. through separate test sets for various types of numeric sequences, or words with unusual grapheme-to-phoneme correspondences, and (4) taking steps to make available any missing linguistic knowledge to the model in the next round of training, or at inference time.

2.1. Semiotic Classes: Numbers, Times, etc.

In our hypothetical example above, '12' would need to be observed in the training data for the model to learn this unusual and arbitrary correspondence. Including it in the training data, then, would be a natural solution. And indeed, our argument is not that an end-to-end model could never learn such edge cases at all—we simply observe, in a flavor of the poverty-of-the-stimulus argument, that the system *must* have observed such idiosyncratic cases at least once to be able to learn them.

Of course, this does raise the question of what needs to be included in the training data. Some correspondences will be generalizable: for example, a system that observed in its training data all the numbers between '20' and '40' apart from '33' should be able to generalize correctly and produce 'thirty-three' for '33', following the pattern of first verbalizing the decade form ('twenty' or 'thirty') and then using the regular cardinal number 'three' (as in

'thirty-one', 'thirty-two', and so on). In other words, some examples are entirely arbitrary and idiosyncratic, and need to be specifically covered individually, while for others, generalization is an option as long as sufficient information is available from which to generalize.

Now, even if a model observes such an entirely idiosyncratic case such as '12' only once at training time, the model may not remember this correspondence; to our knowledge there has been no research determining if there is a threshold of occurrences that is sufficient to teach an end-to-end system a given correspondence. Presumably the degree of arbitrariness and frequency both play a role, as does the general model architecture, but this seems like a rich area for analysis. However, an end-to-end model would at least theoretically stand a chance to get '12' right if it was included even just once in its training data; if it never observed this case at all, it would stand no chance.

If the issues were limited to cases like like '10', '11' and '12', this would perhaps pose only a limited problem that could be easily addressed by including relevant data at training time. However, even for simple cardinal numbers in counting forms like 'one', 'two', 'three', a relatively complex induction needs to be done to derive the correct correspondences even just for forms between 1 and 999, as shown by Ritchie et al. (2019); Gorman and Sproat (2016). These inductions differ in complexity from one language to another, but are rarely straightforward. In the extreme, for some languages, the only option for getting the numbers between 1 and 100 right appears to be explicitly enumerating every single form (Gorman and Sproat, 2016). This would imply that an end-to-end speech system would also need to observe every single form in its training data at least once.

Cardinal numbers are, unfortunately, perhaps among the *easier* correspondences to learn. There are many classes of tokens for which there are non-trivial correspondences between written and spoken forms of human language, and they appear in mostly any language; see e.g. van Esch and Sproat (2017) which provides an overview of these 'semiotic classes', ranging from phone numbers like '1-800-GOOG411' to times like '10:15' ('ten fifteen' or 'a quarter past ten'), and from measures like '10km' to currency amounts like 'HK$300'.

For end-to-end speech models, learning such correspondences is difficult, with low accuracy rates unless special measures are taken (Peyser et al., 2019). In fact, as Sproat and Jaitly (2017); Zhang et al. (2019) show, handling semiotic classes is difficult even for standalone text-to-text neural networks that are dedicated entirely to transforming between spoken-domain text strings like 'twelve' and written-domain text strings like '12'. Sproat and Jaitly (2017); Zhang et al. (2019) point out that even such text-to-text models frequently make so-called 'silly errors' like verbalizing '16GB' into 'sixteen hertz' instead of 'sixteen gigabytes'—even though for these networks, large amounts of relevant data were available at training time, based on which the correct behavior *could* have been inferred.

2.2. Normal Words: Names, Loanwords, Acronyms, etc.

Beyond semiotic-class tokens like '12', '1-800-GOOG411', '10:15', 'HK$300', and '16GB', there are also countless examples of English words with idiosyncratic grapheme-to-phoneme mappings. For example, the pronunciation of the English word 'Worchestershire' is relatively idiosyncratic, consisting of only three syllables. Put simply, the rules of English orthography do not map one-to-one onto English pronunciations—and this is the case in many of the world's languages (though not everywhere). The complexity of various orthographic systems can be measured (van den Bosch et al., 1994), and different orthographic systems are known to have different degrees of orthographic transparency (Katz and Frost, 1992). In practice, this means that in some languages, the correspondences between spoken and written forms will be harder to learn than in others.

Indeed, cross-language comparisons of grapheme-to-phoneme (G2P) conversion models such as (van Esch et al., 2016; Lee et al., 2020) show widely different accuracy rates across languages. While the accuracy metrics achieved by G2P models also depend on the amount of training data available for the target language, as well as factors such as the model architecture, there is unmistakably an impact from the degree of orthographic transparency in each language. Some languages, like Spanish, have a reasonably transparent orthography, and G2P accuracy rates are usually high; languages like English, on the other hand, feature large numbers of idiosyncratic cases, which are much more challenging or even impossible for a G2P model to predict based on the training data, leading to lower accuracy rates.

Such challenging grapheme-to-phoneme correspondences are known to impact the quality of end-to-end speech models: for example, Taylor and Richmond (2019); Fong et al. (2019) show that end-to-end TTS models struggle to generate correct audio for words with irregular or idiosyncratic G2P correspondences. End-to-end ASR systems face similar struggles (Kim et al., 2020; Prabhavalkar et al., 2017), although to our knowledge the issue has not been analyzed in detail. Unusual grapheme-to-phoneme correspondences appear in many types

of words, including in place names like 'Worchestershire', names of people and businesses (Rutherford et al., 2014), names of artists (like 'P!nk', where the '!' stands for an 'i', or 'deadmau5', read as 'dead mouse'), and loanwords (which may retain the original spelling from their source language, as in 'restaurant' or 'La Jolla'). Sometimes, otherwise entirely unremarkable nouns suddenly involve an unpredictable correspondence, as in 'sword', the only word in the English language where the grapheme 'w' is silent in the onset cluster 'sw': compare, for example, 'swam', 'sweep', and 'swore'. Highly frequent words may also have unusual grapheme-to-phoneme correspondences, like English 'one'. And letter sequences (Sproat and Hall, 2014) such as 'NASA' (read as a word) and 'C-SPAN' (partially read as a word, partially as a letter) present their own idiosyncrasies—not to mention borrowed letter sequences, such as 'BBC', which is read letter-by-letter using English letter pronunciations in many European languages.

As with semiotic-class tokens, it can range from challenging to impossible for an end-to-end speech model to predict the correct grapheme-to-phoneme correspondence for a given word, depending on the degree of arbitrariness of the relationship, the training data, and the model's generalization ability. According to the Census Bureau, there are tens of thousands of geographical names in the United States alone. Many of them are likely reasonably amenable to generalization, but others (like 'La Jolla') will not be, and must be observed in the training data for an end-to-end model to learn them. In the extreme, cases like 'deadmau5' are presumably sufficiently idiosyncratic as to be impossible for an end-to-end system to predict correctly through generalization from any other training data.

2.3. Homophones, Homographs, and Context

For both semiotic-class tokens and normal words, another issue can cause further challenges for ASR, namely homophony—words or phrases that sound the same, but have different spellings depending on their meaning and context. For example, 'three eleven' could be written as '3:11' (as a time) or '3/11' (as a date), and 'Xanh' (a popular restaurant in Mountain View, California, which unfortunately closed after the pandemic) shares its pronunciation with 'sun'. As an extreme example, if these terms were only ever observed in isolation at training time, the system would find it challenging to determine that it should emit 'dinner at Xanh in Mountain View' (not 'sun') but 'the sun is shining' (not 'Xanh').

In TTS applications, homographs, or words that are spelled the same but have different pronunciations depending on context, pose similar problems: 'Houston, Texas' is pronounced differently than 'Houston Street, New York City' (which is pronounced like 'how-ston' not 'hew-ston'), but again, if the term 'Houston' was only ever observed in isolation, the model would struggle to decide which of the two pronunciations to use based on inference-time context—assuming, of course, that both pronunciations were even included in the training data.

3. Mitigation Techniques

The existence of such arbitrariness is not an argument against end-to-end modeling: our goal has only been to point out that it is impossible for an end-to-end model to correctly predict an entirely arbitrary phenomenon that it has not observed at training time—and that even for slightly less arbitrary phenomena, such models may struggle. But given the benefits of end-to-end modeling, it is clearly desirable to see if these challenges can be mitigated within the end-to-end paradigm.

Before discussing mitigation strategies, one question that may come to mind is whether any mitigation is in fact needed at all: one might argue that handling these correspondences can simply be called out-of-scope entirely. For example, an ASR system could simply emit 'twelve' instead of '12', and a TTS system could simply require that only forms like 'twelve' are used in any input text. However, in a real-world system this is typically infeasible or impractical, given that TTS applications are generally expected to be able to handle generic written-domain text, and given that downstream processing of ASR transcriptions generally also relies on written forms like '11:15'—for example, in conversational voice assistants that need to identify times in transcribed spoken commands. Taking this position is even harder in the case of words with unusual grapheme-to-phoneme relationships: it would be hard to argue that general-purpose ASR or TTS systems do not need to correctly pronounce or transcribe phrases like 'La Jolla'.

First, we recommend setting up evaluation sets that specifically aim to measure ASR or TTS quality for different categories of arbitrary correspondences, such as words with unusual grapheme-to-phoneme relationships, and different types of semiotic classes, as in Peyser et al. (2019). Such sets will help us understand the extent to which these problems appear for a given model. The question then becomes how to maximize accuracy for these cases.

3.1. Large Training Data Sets

Ensuring that end-to-end systems see sufficient data to correctly generalize all generalizable correspondences, and to learn even the most arbitrary

cases, is one possible approach. This does pose some practical problems, since there will be *many* words that are affected (especially in languages like English, with its opaque orthography). But as one increases the size of the training data, more correspondences will be covered, mitigating the problem to some extent—and with the abundance of data in high-resource languages, the problem may even be invisible entirely unless specific evaluations are done, as in Peyser et al. (2019).

Theoretically, one could simply collect recordings of *all* normal words and semiotic-class tokens in the target language, but this would clearly be very time-consuming, and it is not clear that it poses less of a bottleneck than creating verbalization grammars and phonemic dictionaries. In the extreme, it is arguably impossible due to the infinite amount of e.g. cardinal numbers, but at at any rate, recording many hundreds of thousands words and phrases would be challenging even for ASR, where training data can be gathered from many speakers through platforms such as Hughes et al. (2010); for TTS, where high-quality single-speaker recordings have typically been required, it would be entirely impractical.

In addition to practical factors that make adding more training data a less-than-desirable mitigation strategy, recent work also suggests that reasonable levels of ASR or TTS quality can be obtained by using only an hour or less of supervised target-language audio (Baevski et al., 2020; He et al., 2021), combined with self-supervised learning techniques and/or multilingual modeling. Such work is incredibly promising for addressing the single biggest bottleneck in bringing speech technologies to more languages, namely the scarcity of supervised training data, but it seems vanishingly unlikely that 40 seconds of target-language audio (as in He et al. (2021)) could contain sufficient information to learn all relevant arbitrary correspondences for the target language.

In some cases, multilingual modeling may help, e.g. in predicting that '24' should be verbalized as 'vingt-quatre' in French, following the English pattern. But equally, multilingual modeling may be ineffective or even harmful for this problem, e.g. when mixing English with German, where the correct verbalization of '24' is not 'zwanzig-vier' (literally 'twenty-four'), but 'vier-und-zwanzig' ('four-and-twenty').

3.2. Supplementing Training Data with Synthetic Audio

For ASR, another technique is to use TTS to generate synthetic data to supplement the training data (Rosenberg et al., 2019): for example, Peyser et al. (2019) showed that if a target-language TTS system is available, it can be used to generate transcribed-audio training examples for cases like '12' and 'twelve' at very large scale. However, such approaches still requires the creation of some kind of verbalization grammar (Sak et al., 2013; Ritchie et al., 2019, 2020) to provide the correspondences between the written-domain forms (like '12') which would serve as the ASR training target, and the spoken-domain forms (like 'twelve') which would be passed into the TTS system. And unless the target-language orthography is extremely transparent, the TTS system itself will likely require a phonemic dictionary (recall cases like 'one') in order for the synthetic audio it generates to have the correct pronunciation. Similar synthetic-audio approaches can be employed for phrases like 'La Jolla' with challenging grapheme-to-phoneme correspondences, but again this would require a phonemic dictionary to drive the generation of accurate synthetic audio. In other words, such synthetic-data approaches still require an investment in linguistic resources that is no different from the investments needed to build the linguistic components of conventional, non-end-to-end systems.

3.3. Secondary Models

Yet another class of mitigation techniques involves combining secondary models with the original end-to-end model. For example, fusion techniques are commonly used to connect end-to-end ASR models with neural text-only language models to cover phrases that were not observed in the original training data (Kim et al., 2020). While it seems reasonable that external language models can help with contextual disambiguation ('Xanh' vs 'sun'), their effect for words with unusual grapheme-to-phoneme or arbitrary verbalization correspondences is unclear and requires further research; they are unlikely to be a panacea, especially for highly idiosyncratic cases. In another example, Serrino et al. (2019) describes a module that allows for the use of phonemic dictionaries to correct misrecognitions from the upstream ASR system, but again at the cost of requiring linguistic resources. In both ASR and TTS, secondary neural models can also be used for normalizing semiotic-class tokens before or after the core end-to-end model, as in Zhang et al. (2019); Peyser et al. (2019); such models do, however, typically require large amounts of text-to-text training data, as well as covering grammars, both of which again require linguistic expertise.

3.4. Combining End-to-End and Conventional Approaches

It is also possible to combine the best of both worlds, so to speak, by training models using the end-to-end paradigm which do however still use phonemes

as an input or output unit: one recent example of this is the Hybrid Autoregressive Transducer (HAT) (Variani et al., 2020) for ASR, which combines end-to-end models that output phoneme units with a traditional finite-state transduction decoding graph that uses a phonemic dictionary and verbalization grammars. Similarly, in TTS, end-to-end models can simply take phonemes produced by a conventional text normalization front-end as input, e.g. as in Skerry-Ryan et al. (2018); Yasuda et al. (2020). In a related approach, (Kastner et al., 2019) describes an end-to-end TTS model that allows the mixing of graphemes and phonemes in inference-time inputs, allowing per-example control through phonemic specifications where needed—but then the question becomes how to decide when such control should be exercised. Such best-of-both-worlds approaches share one commonality: they still require the same linguistic components as non-end-to-end systems.

4. Conclusions

End-to-end speech models face challenges when it comes to handling words with unusual grapheme-to-phoneme correspondences (e.g. place names and loanwords) and semiotic classes (e.g. numbers and time expressions), because there is a large amount of arbitrariness in the correspondences between spoken and written forms of human language, and because training data suffers from poverty-of-the-stimulus issues. Traditional speech systems solved these challenges through the explicit injection of linguistic knowledge, e.g. through phonemic dictionaries or verbalization grammars. With thousands of languages spoken in our world, and very few of them covered by speech technologies today, it would be great if we did not need to curate such resources for every language, but this is unlikely to be possible, given that the mitigation strategies we reviewed still require similar amounts of linguistic resources, as we have discussed.

Importantly, we called out that the mitigation strategy employed (mostly implicitly) for many end-to-end systems will no longer work as end-to-end approaches take hold that use relatively small amounts of supervised training data. These approaches rely heavily on self-supervised learning and multilingual modeling—an exciting development that promises to help bring speech technologies to many more languages. At the same time, as we have seen, these methods will likely need to be combined with synthetic-data approaches (in ASR), or best-of-both-world architectures, like the use of phonemes produced by a conventional text normalization front-end as the input unit to end-to-end TTS models, or like HAT (Variani et al., 2020) in ASR.

Fortunately, much work has been done to make creating linguistic resources like phonemic dictionaries and verbalization grammars for new languages easier than ever (Kim and Snyder, 2013; Rutherford et al., 2014; Deri and Knight, 2016; Lee et al., 2020; Ritchie et al., 2019; Bleyan et al., 2019; Ritchie et al., 2020), leading us to be optimistic about the opportunities for bringing high-quality ASR and TTS systems to more languages by combining conventional linguistic resources with innovative modeling approaches that require little supervised audio training data.

5. Acknowledgements

Many thanks to Richard Sproat, Trevor Strohman, Tara Sainath, Jonas Fromseier Mortensen, Pedro Moreno, and Jeremy O'Brien for many lively water cooler conversations on this topic over the years, and for their thoughtful feedback.

6. Bibliographical References

Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. Multilingual acoustic and language modeling for ethio-semitic languages. In *Proceedings of Interspeech 2020*, pages 1047–1051.

Oliver Adams et al. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of NAACL 2019*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

Oliver Adams et al. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. In *Proceedings of ComputEL-4*.

Adam Albright. 2009. Lexical and morphological conditioning of paradigm gaps. *Modeling ungrammaticality in optimality theory*, pages 117–164.

Rosana Ardila et al. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of LREC 2020*, pages 4218–4222, Marseille, France. ELRA.

Alexei Baevski et al. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv:2006.11477*.

Laurent Besacier et al. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communications*, 56:85–100.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.*, 50(5):434–451.

David Blachon et al. 2016. Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app. In *Proceedings of SLTU 2016*, Yogyakarta, Indonesia.

Harry Bleyan et al. 2019. Developing pronunciation models in new languages faster by exploiting common grapheme-to-phoneme correspondences across languages. In *Proceedings of Interspeech 2019*.

Nicholas Buckeridge and Ben Foley. 2020. Scaling language data import/export with a data transformer interface. In *Proceedings of SLTU-CCURL 2020*, pages 121–125, Marseille, France. ELRA.

Isaac Caswell et al. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of COLING 2020*.

Isaac Caswell et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv:2103.12028*.

Tania Chakraborty et al. 2021. A large scale low-resource pronunciation data set mined from Wikipedia. *arXiv:2101.11575*.

Po-Han Chi et al. 2021. Audio ALBERT: A lite BERT for self-supervised learning of audio representation. *arXiv:2005.08575*.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2017. State-of-the-art speech recognition with sequence-to-sequence models. *CoRR*, abs/1712.01769.

Mason Chua et al. 2018. Text normalization infrastructure that scales to hundreds of language varieties. In *Proceedings of LREC 2018*.

Yu-An Chung et al. 2018. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv:2006.13979*.

J. Cui et al. 2015. Multilingual representations for low resource speech recognition and keyword search. In *ASRU 2015*, pages 259–266.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of ACL 2016*, pages 399–408, Berlin, Germany. ACL.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

Siyuan Feng et al. 2021. How phonotactics affect multilingual and zero-shot ASR performance. In *ICASSP 2021*.

Ben Foley et al. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system. In *Proceedings of SLTU 2018*.

Jason Fong, Jason Taylor, Korin Richmond, and Simon King. 2019. A comparison of letters and phones as input to sequence-to-sequence models for speech synthesis. In *Proceedings of ISCA SSW 2019*, pages 223–227.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Anmol Gulati et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040.

Vishwa Gupta and Gilles Boulianne. 2020. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of LREC 2020*, pages 2521–2527, Marseille, France. ELRA.

Mark Hasegawa-Johnson, Camille Goudeseune, and Gina-Anne Levow. 2019. Fast transcription of speech in low-resource languages. *arXiv:1909.07285*.

Mark Hasegawa-Johnson et al. 2020. Grapheme-to-phoneme transduction for cross-language ASR. In *Statistical Language and Speech Processing*, pages 3–19, Cham. Springer International Publishing.

Tomoki Hayashi et al. 2020. Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *ICASSP 2020*, pages 7654–7658. IEEE.

Mutian He, Jingzhou Yang, and Lei He. 2021. Multilingual byte2speech text-to-speech models are few-shot spoken language learners. *arXiv:2103.03541*.

Stephanie Hirmer et al. 2021. Building representative corpora from illiterate communities: A review of challenges and mitigation strategies for developing countries. In *Proceedings of EACL 2021*.

Nils Hjortnaes et al. 2020. Improving the language model for low-resource ASR with online text corpora. In *Proceedings of SLTU-CCURL 2020*, pages 336–341, Marseille, France. ELRA.

Mathieu Hu et al. 2020. Kaldi-web: An installation-free, on-device speech recognition system. In *Proceedings of Interspeech 2020: Show & Tell*, Shanghai, China.

Thad Hughes et al. 2010. Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of Interspeech 2010*, pages 1914–1917.

Pratik Joshi et al. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020*, pages 6282–6293, Online. ACL.

Anjuli Kannan et al. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proceedings of Interspeech 2019*, pages 2130–2134.

Kyle Kastner, João Felipe Santos, Yoshua Bengio, and Aaron Courville. 2019. Representation mixing for TTS synthesis. In *Proceedings of ICASSP 2019*.

Leonard Katz and Ram Frost. 1992. The reading process is different for different orthographies: The orthographic depth hypothesis. In Leonard Katz and Ram Frost, editors, *Orthography, Phonology, Morphology, and Meaning*, page 67–84. Elsevier North Holland Press, Amsterdam.

Chanwoo Kim, Dhananjaya Gowda, Dongsoo Lee, Jiyeon Kim, Ankur Kumar, Sungsoo Kim, Abhinav Garg, and Changwoo Han. 2020. A review of on-device fully neural end-to-end automatic speech recognition algorithms. *arXiv:2012.07974*.

Young-Bum Kim and Benjamin Snyder. 2013. Optimal data set selection: An application to grapheme-to-phoneme conversion. In *Proceedings of NAACL 2013*, pages 1196–1205, Atlanta, Georgia. ACL.

S. H. Krishnan Parthasarathi and N. Strom. 2019. Lessons from building acoustic models with a million hours of speech. In *ICASSP 2019*, pages 6670–6674.

Stephen Laurence and Eric Margolis. 2001. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, 52(2):217–276.

Jackson L. Lee et al. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of LREC 2020*, pages 4223–4228.

B. Li, Y. Zhang, T. N. Sainath, Y. Wu, and W. Chan. 2019. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *Proceedings of ICASSP 2019*.

Xinjian Li et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020*, pages 8249–8253. IEEE.

Yusen Lin, Jiayong Lin, Shuaicheng Zhang, and Haoying Dai. 2021. Bilingual dictionary-based language model pretraining for neural machine translation.

Chunxi Liu et al. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of SLTU-CCURL 2020*, pages 46–52, Marseille, France. ELRA.

M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of LREC 2018*, Miyazaki, Japan. ELRA.

A. Oktem et al. 2020. Gamayun - language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference*, pages 1–4.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of ACL 2020*, pages 1703–1714, Online. ACL.

V. Panayotov et al. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP 2015*, pages 5206–5210.

Daniel S. Park et al. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. *Proceedings of Interspeech 2019*, pages 2613–2617.

Daniel S. Park et al. 2020. SpecAugment on large scale datasets. *ICASSP 2020*, pages 6879–6883.

Niko Partanen and Michael Rießler. 2019. An OCR system for the Unified Northern Alphabet. In *Proceedings of the Fifth Workshop on Computational Linguistics for Uralic Languages*, pages 77–89, United States. ACL.

Matthias Petursson, Simon Klüpfel, and Jon Gudnason. 2016. Eyra - speech data acquisition system for many languages. In *Proceedings of SLTU 2016*, Yogyakarta, Indonesia.

Cal Peyser et al. 2019. Improving performance of end-to-end ASR on numeric sequences. In *Proceedings of Interspeech 2019*.

Jonas Pfeiffer et al. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of EMNLP 2020: Systems Demonstrations*, pages 46–54, Online. ACL.

Daniel Povey et al. 2011. The Kaldi speech recognition toolkit. In *ASRU 2011*.

Rohit Prabhavalkar, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A comparison of sequence-to-sequence models for speech recognition. In *Proceedings of Interspeech 2017*, pages 939–943.

Manasa Prasad, Theresa Breiner, and Daan van Esch. 2018. Mining training data for language modeling across the world's languages. In *Proceedings of SLTU 2018*.

Manasa Prasad et al. 2019. Building large-vocabulary asr systems for languages without any audio training data. In *Proceedings of Interspeech 2019*.

Vineel Pratap et al. 2020. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Proceedings of Interspeech 2020*, pages 4751–4755.

S. Punjabi, H. Arsikere, and S. Garimella. 2019. Language model bootstrapping using neural machine translation for conversational speech recognition. In *ASRU 2019*, pages 487–493.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR post correction for endangered language texts. In *Proceedings of EMNLP 2020*, pages 5931–5942, Online. Association for Computational Linguistics.

Sandy Ritchie et al. 2019. Unified verbalization for speech recognition synthesis across languages. In *Proceedings of Interspeech 2019*.

Sandy Ritchie et al. 2020. Data-driven parametric text normalization: Rapidly scaling finite-state transduction verbalizers to new languages. In *Proceedings of SLTU-CCURL 2020*.

Andrew Rosenberg et al. 2019. Speech recognition with augmented synthesized speech. In *ASRU 2019*.

Attapol Rutherford, Fuchun Peng, and Françoise Beaufays. 2014. Pronunciation learning for named-entities through crowd-sourcing. In *Proceedings of Interspeech 2014*.

Tara N. Sainath, Rohit Prabhavalkar, Shankar Kumar, Seungji Lee, Anjuli Kannan, David Rybach, Vlad Schogol, Patrick Nguyen, Bo Li, Yonghui Wu, Zhifeng Chen, and Chung-Cheng Chiu. 2018. No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models. In *Proceedings of ICASSP 2018*.

H. Sak et al. 2013. Language model verbalization for automatic speech recognition. In *ICASSP 2013*.

O. Scharenborg et al. 2020. Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:964–975.

S. Schneider et al. 2019. wav2vec: Unsupervised pre-training for speech recognition. *Proceedings of Interspeech*, pages 3465–3469.

Tanja Schultz and Alex Waibel. 2001. Experiments on cross-language acoustic modeling. In *Proceedings of the 7th European Conference on Speech Communication and Technology*.

Frank Seifart et al. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Jack Serrino, Leonid Velikovich, Petar Aleksic, and Cyril Allauzen. 2019. Contextual recovery of out-of-lattice named entities in automatic speech recognition. In *Proceedings of Interspeech 2019*, pages 3830–3834, Graz, Austria.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end asr for endangered language documentation: An empirical study on yoloxóchitl mixtec.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *Proceedings of ICML 2018*.

Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2Wav: End-to-end speech synthesis. In *Proceedings of ICLR 2017*.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

Richard Sproat and Keith Hall. 2014. Applications of maximum entropy rankers to problems in spoken language processing. In *Proceedings of Interspeech 2014*.

Richard Sproat and Navdeep Jaitly. 2017. RNN approaches to text normalization: A challenge. *arXiv:1611.00068*.

Piotr Szymański et al. 2020. WER we are and WER we think we are. In *Findings of EMNLP 2020*, pages 3290–3295, Online. ACL.

Jason Taylor and Korin Richmond. 2019. Analysis of Pronunciation Learning in End-to-End Speech Synthesis. In *Proceedings of Interspeech 2019*, pages 2070–2074.

Anjana Vakil et al. 2014. lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition. In *Proceedings of ACL 2014: System Demonstrations*, pages 109–114, Baltimore, Maryland. ACL.

A.P.J. van den Bosch, A. Content, W.M.P. Daelemans, and B.L.M.F. de Gelder. 1994. Measuring the complexity of writing systems. *Journal of Quantitative Linguistics*, 1(3):178–188. Pagination: 11.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Daan van Esch, Mason Chua, and Kanishka Rao. 2016. Predicting pronunciations with syllabification and stress with recurrent neural networks. In *Proceedings of Interspeech 2016*.

Daan van Esch and Richard Sproat. 2017. An expanded taxonomy of semiotic classes for text normalization. In *Proceedings of Interspeech 2017*.

Daan van Esch et al. 2019. Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard. Technical report.

Nanne van Noord et al. 2021. Automatic annotations and enrichments for audiovisual archives. In *ICAART 2021*.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (HAT). In *Proceedings of ICASSP 2020*.

Shafqat Mumtaz Virk et al. 2020. The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of LREC 2020*, pages 878–884, Marseille, France. ELRA.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech 2017*, pages 4006–4010.

Shinji Watanabe et al. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech 2018*, pages 2207–2211.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of SLTU-CCURL 2020*, pages 306–315, Marseille, France. ELRA.

Yusuke Yasuda, Xin Wang, and Junichi Yamagishi. 2020. Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis. *arXiv:2005.10390*.

Piotr Zelasko et al. 2020. That Sounds Familiar: An Analysis of Phonetic Representations Transfer Across Languages. In *Proceedings of Interspeech 2020*, pages 3705–3709.

Hao Zhang, Richard Sproat, Axel H. Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Comput. Linguist.*, 45(2):293–337.