# Intro to the IARPA Babel corpora and discussion of implications for ASR/ST/Grapheme-Phoneme tasks

P. R. Sullivan[1]

[1]School of Information
UBC

UBC DL-NLP, September 2020

# Table of Contents

# Objective

I'm looking at this as a discussion of a set of corpora that I am planning on incorporating in my thesis, that seem like they may be useful for other tasks. To that end, I'll be giving an overview of my current Thesis plan, look at relevant papers that have been using Babel, as well as discuss how it might be used in other work (especially Grapheme-Phoneme models).

# Table of Contents

# Thesis topic

- Low-resource ST is often dependant on pre-training on other languages.
- I originally wanted to explore comparing pre-training on a set of languages in the same family (and to pick very closely related languages to do this).
- Recent work did this on Indo-European (pre-train on Portugese or French and test on Spanish), and showed that language family wasn't so important as WER of ASR module. Further showed that you could train on unrelated languages with good results (Chinese!!) [1]

# Thesis topic cont.

- If language family isn't important, is this true for phonological closeness as well?
- Recent work also shows that Phones can be extremely important for ST [2]

Which leads to a number of questions:

- Could it be possible to define some metric for phonological closeness between languages?
- Does this even matter at the language level, could we focus instead on utterances-level to train on phonologically close sentences regardless of language origin?
- How much does phonological composition matter? (Chinese and Spanish both have relatively simple vowels and thus individual phoneme overlap, but longer range dependencies are quite different)

# Table of Contents

# Refresher on types of Speech corpora

- Generally speaking, most ASR (and to a degree ST) corpora simply map from audio to a grapheme-based transcript (e.g. written transcript of the dialogue in the native writing system of that language).
- However, earlier HMM-GMM/HMM-DNN approaches tended to use TIMIT style corpora that were hand annotated based on the actual pronunciation of the speakers in the dialogues. This proved to be expensive and error prone.
- A third approach, exemplified by Babel (and say CALLHOME another set of corpora), is to simply provide a lexicon that gives pronunciations of words in the corpus (including variant pronunciations that might occur).

# Quick Look

Type 1. Standard orthography used as target labels. Example:

```
0 46797 She had your dark suit in greasy wash water all year.
```

# Quick Look

## Type 2. TIMIT Transcript + Phone-Transcription Example:

```
0 46797 She had your dark suit in greasy wash water all year.
```

```
0 3050 h#
3050 4559 sh
4559 5723 ix
5723 6642 hv
6642 8772 eh
8772 9190 dcl
9190 10337 jh
10337 11517 ih
11517 12500 dcl
12500 12640 d
12640 14714 ah
14714 15870 kcl
```

# Quick Look

Type 3. Transcript With an additional Lexicon.

Example from IARPA Babel Lithuanian TRANSCRIPT:

```
[22.055]
tai gerai nes biškį (()) per atstumą čia tokioj apklausoj bišl
[27.735]
<no-speech>
[28.585]
<int> nu kaip <int> <no-speech> kaip gyveni
[32.515]
<no-speech>
[33.745]
<sta> nu
[34.265]
<no-speech>
[35.835]
gerai aš žinok <hes> aš naują darbą turiu <no-speech> ką
```

# Quick Look

Type 3. Transcript With an additional Lexicon.
Example from IARPA Babel English Lexicon:

| WORD | | PHONEMIC TRANSCRIPTIONS | | | | |
|---|---|---|---|---|---|---|
| chin | \t | " tS I n | | | | |
| cut | \t | k V t | | | | |
| either | \t | " aI . D @ | \t | " i: . D @ | | |
| heard | \t | "h e r d | \t | "h 3: d | \t | "3: d |
| mock | \t | " m Q k | | | | |
| pin | \t | " p I n | | | | |
| read | \t | " r i: d | \t | " r E d | | |
| red | \t | " r E d | | | | |
| thin | \t | " T I n | | | | |

# Notes on Babel

26 Languages, each of about 215 hours of recorded audio, half of which is transcribed, half is raw audio.

- Cantonese
- Assamese
- Bengali
- Pashto
- Turkish
- Georgian
- Tagalog
- Vietnamese
- Haitian
- Creole
- Swahili
- Zulu
- Kurmanji
- Kurdish
- Tok Pisin
- Cebuano
- Kazak
- Telugu
- Lithuanian
- Guarani
- Igbo
- Amharic

# Table of Contents

# Papers using Babel

- [3] Finds that Multitask learning (with ST as aux task) while performing ASR-pretraining allows for improved performance on final ST. Uses two Babel languages for target ASR.
- [4] Compares Multitask Learning pre-training vs. MetaLearning pretraining (MAML) to improve low-resource E2E ASR. Uses a set of Babel languages.
- [5] Investigates Language Model fusion for improving multilingual ASR finetuning. Uses a set of 10 Babel languages to pre-train on.
- [6] Introduces Epitran a Grapheme to Phoneme tool for 60+ languages. This uses Babel as a validation set.

# Table of Contents

# Questions for discussion

- Can we see these corpora potentially be used for other tasks? For instance phone features to improve text MT?
- What might be issues with sampling from such a diverse set of languages?
- What might be ways to better measure overlap between languages? Syntactically? Phonetically?

# References I

[1] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7909–7913.

[2] E. Salesky and A. W. Black, "Phone features improve speech translation," *arXiv preprint arXiv:2005.13681*, 2020.

[3] C. Wang, J. Pino, and J. Gu, "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation," *arXiv preprint arXiv:2006.05474*, 2020.

# References II

[4] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7844–7848.

[5] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end asr with language model fusion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6096–6100.

# References III

[6] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision g2p for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.