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

Building Large-Vocabulary ASR Systems for Languages Without Any Audio Training Data

Manasa Prasad, Daan van Esch, Sandy Ritchie, Jonas Fromseier Mortensen
Google
{pbmanasa, dvanesch, sandyritchie, jfmortensen}@google.com

Abstract

When building automatic speech recognition (ASR) systems, typically some amount of audio and text data in the target language is needed. While text data can be obtained relatively easily across many languages, transcribed audio data is challenging to obtain. This presents a barrier to making voice technologies available in more languages of the world. In this paper, we present a way to build an ASR system system for a language even in the absence of any audio training data in that language at all. We do this by simply re-using an existing acoustic model from a phonologically similar language, without any kind of modification or adaptation towards the target language. The basic insight is that, if two languages are sufficiently similar in terms of their phonological system, an acoustic model should hold up relatively well when used for another language. We describe how we tailor our pronunciation models to enable such re-use, and show experimental results across a number of languages from various language families. We also provide a theoretical analysis of situations in which this approach is likely to work. Our results show that it is possible to achieve less than 20% word error rate (WER) using this method.

Index Terms: zero shot speech recognition, low resource speech recognition

1. Introduction

Thousands of languages are spoken in our world today [1], but technologies such as automatic speech recognition (ASR) are not yet available in the vast majority of these languages. A full count is hard to obtain, but commercial APIs typically support at most about 100 different language varieties. ASR systems have been developed for more languages, typically by academic researchers, who may be working with multilingual data sets like IARPA Babel [2], or who may be using only a monolingual data set in their target language.

Being able to easily create ASR systems for more languages would be tremendously helpful for language communities around the world, especially as smartphone penetration has grown [3, 4]. For example, ASR systems can help low-literacy users access information by enabling them to search by voice [5], and they can make it easier to communicate by allowing for voice dictation of messages [6].

1.1. High development costs

The unavailability of ASR systems in the vast majority of the world's languages is frequently explained by pointing to the high development costs per language, along with the fact that language sizes are distributed unevenly, such that even supporting only the top 10 biggest languages by number of speakers already covers more than half of the world's population [7].

It follows that to build ASR systems more easily across languages, it would be useful to focus on decreasing the development costs. To achieve this, it helps to understand the development costs in some more detail. Typically, ASR systems developed in the finite-state transduction framework [8] require a large transcribed audio corpus to train an acoustic model (typically at least on the order of hundreds of hours); a pronunciation model or lexicon which provides pronunciations for all the words in the system; a text corpus to train a language model; and a smaller audio corpus to test the entire system (usually a few thousand recordings).

Of these resources, audio training corpora tend to require the most effort to obtain, despite advanced crowd-sourcing tools and platforms like DataHound [9], Aikuma [10] and Common Voice [11]. After all, building these audio corpora requires speakers to record hundreds or even thousands of hours of speech (and to verify the transcriptions). On the other hand, text corpora can be crawled from the open web [12] and cleaned automatically [12]. Development of pronunciation models can also be done in a relatively straightforward way for most languages, as long as their orthographies are sufficiently transparent [13, 14, 15]. If needed, grammars for handling numeric entities can also be created relatively easily [16, 17].

1.2. Lowering development costs

To reduce development costs as much as possible, ideally, we would focus on reducing the size of the audio training data sets needed to build an ASR system for a new language. Naturally, this is not a novel idea: significant amounts of research have gone into building ASR systems for languages with only small amounts of audio data available (so-called "low-resource" languages). Typically, such research uses techniques like multilingual modeling, adaptation, and transfer learning; see e.g. [18, 19, 20]. In some cases, such systems are built using unsupervised audio training data, e.g. in [21, 22].

While many of these papers show promising results, very few (if any) completely avoid the need for any audio training data in the target language; one recent system is [23] but this system produces only phoneme-level transcriptions, and has relatively high phoneme error rates. Given the basic linguistic insight that many closely related languages share similar phonological systems [24], we wondered if it would be possible to simply re-use a previously-existing acoustic model (AM) from another language, without any modifications to the AM whatsoever. Doing so would reduce the resources needed for development of a new language significantly, by avoiding the need for an acoustic training data corpus entirely. We would still need a pronunciation model and a language model (LM) in the target language to build a decoding graph, but as mentioned above, these resources are relatively easy to acquire.

http://dx.doi.org/10.21437/Interspeech.2019-1775

2. Selecting Languages

To give an informal definition, we expected our approach to work for any target language with a phonological system that is sufficiently similar to another language with a previously-existing AM (the "source" language), as long as this target language also has a sufficient amount of text to train an LM.

This is somewhat unsatisfying: after all, when are two languages sufficiently similar? While a number of databases with phonemic inventories are now available [25, 26], it was not clear how we might automatically identify promising candidate pairs, on the assumption that we are given a list of source languages with existing AMs. We decided to leave the creation of such automatic methods to future work: the future-work section contains our thoughts around creating a more systematic approach. For now, we asked in-house linguists to identify any candidates among the list of about 400 languages with more than 1 million speakers according to the Ethnologue [1]. We restricted ourselves to these relatively large languages since we needed to identify candidates manually, but our approach should work for any language, as long as some sufficiently phonologically similar high-resource language can be identified and as long as a text corpus for the target language can be found. Future work on automatic candidate identification would enable analyses to be done for the world's remaining thousands of languages.

Our in-house linguists identified a few dozen potential candidates, and we selected four experimental languages from three different language families (see table 1). For rapid testing, we also identified some candidate languages that we had already built a full ASR system for, including an AM trained on a large audio training data corpus, so we could do ablation experiments by simply pretending we did not have this audio training data. Even if these languages are phonologically similar to an already-supported language, they differ in terms of their G2P relationship, their vocabulary, and their grammatical structures.

3. Experiments

3.1. Ablation experiments vs. full-blown acoustic models

To test our theory, we chose two Indian languages we have already built and launched an ASR system for, namely Marathi and Gujarati. Both languages are phonologically similar to Hindi in our assessment, so we used our Hindi AM. The Hindi AM is a CD-CTC-SMBR [27] model trained on an anonymized hand-transcribed corpus of 18K hours, representative of Google's traffic. Two anonymized hand-transcribed corpora of about 10 hours each serve as our test sets for Marathi and Gujarati, again representative of our traffic in these languages.

Our baseline Word Error Rate (WER) for Marathi was about 50.5%, where our set-up was using an CD-CTC-SMBR AM trained on about 230 hours of Marathi read speech, combined with a set of grapheme-to-phoneme (G2P) conversion rules created for Marathi by linguistic experts, and a 5-gram language model (LM) trained consisting of 15M Marathi ngrams overall, trained on text mined using the approaches in [12]. In our experiment, we replaced this Marathi AM with our Hindi AM. Our linguistic team also modified our Marathi G2P rules to use only those phonemes which the Hindi AM would recognize, taking our phonemic transcriptions in Marathi and mapping them to the nearest possible Hindi phonemic transcriptions based on phonetic and phonological similarity.

This set-up for Marathi ASR (using the Hindi AM, a set of Marathi-to-Hindi G2P rules, and the same n-gram LM for Marathi) produced a WER of 47.4%, i.e. slightly outperforming our baseline, which used an AM trained specifically on Marathi audio data. We theorize that this is in part because of the larger size of the training corpus for the Hindi AM, and in part because of the spontaneous-speech nature of the Hindi training data, which matches our Marathi test set. By contrast, our Marathi AM was trained only on read speech.

We tried the same experiment on Gujarati, and obtained similar results: our baseline WER here was 74.8%, where the CD-CTC-SMBR AM was again trained on Gujarati read speech (in this case, about 370 hours), and where we used Gujarati-specific G2P rules as well as a 5-gram LM trained on 15M ngrams of Gujarati text. Swapping out the Gujarati AM with the Hindi AM described above, and making similar modifications to our Gujarati G2P rules, yielded a word error rate of 71.2%, again slightly outperforming the baseline WER.

Given these results, we experimented to see what would happen if we simply evaluated our Marathi and Gujarati test sets using the regular Hindi recognizer (i.e. using our Hindi AM, a regular Hindi pronunciation model, and a Hindi LM). For our Marathi test set, this yields 60.2% WER; our Gujarati test set yields 102.8% WER. The high WER on Gujarati is expected, as it uses a different script and there is basically no overlap in the vocabulary of the Hindi recognizer and the words used in the Gujarati test set. For Marathi, some vocabulary overlap exists, but we still see significantly higher accuracy when using pronunciation and LMs from the target language, with some mappings applied to the pronunciation models. Our approach is even more effective in situations such as in Gujarati, where there is little or no overlap in vocabulary between the source-language recognizer and the target language.

3.2. New languages

Once we were satisfied that our approach seemed promising, we decided to move forward with some candidate languages for which we had not yet built any ASR systems. We gathered small audio data sets to test the quality of the resulting system during our development process. In practice, this is not strictly a requirement for adopting this approach: it would be possible to simply ask some speakers of the target language to try out the system once it has been built, which may be easier than recording a test set separately, and which would further lower the development costs for a new language. However, for our purposes, we worked with test sets to enable quick development and iteration. Specifically, we worked on Cebuano, Kyrgyz, Corsican, and Maithili. We asked participants to read a set of about 10K sentences to create a test set for each of these languages.

To build LMs in these target language varieties, we used the text mining and normalization approaches described in [12, 28]. We built n-gram models based on the text corpora we were able to mine, and where available, added in additional unigrams based on wordlists for additional LM coverage. We used G2P rules and basic verbalizers, specifically created by our linguistic team for each of the target languages. As in the ablation experiments, our linguistic team also created phoneme mappings between the phonemes recognized by the source-language AM and the phoneme inventory of the target language. The AMs were again CD-CTC-SMBR models trained on anonymized hand-transcribed corpora, representative of Google's traffic. Specifically, we used models from Turkish (trained on 15K hours), Italian (trained on 11.6K hours), and Filipino (trained on 1K hours), as well as the Hindi model we used in our experiments above. For each target language, we also calculated the baseline WER by simply feeding the

target-language test set into the full ASR system of the source language. Where the phoneme mappings were not clearly 1:1 or where the target language had phonemes not present in the source language, we experimented with different mappings. We also experimented with different source languages for comparison, for example Kyrgyz using Swedish and Turkish AMs as the source. The best results are presented in table 1.

3.2.1. Analysis

We see that Cebuano and Kyrgyz perform quite well, achieving around 20% WER. In the case of Kyrgyz, the baseline WER is to be expected because it uses a different script than Turkish, similar to the Gujarati/Hindi experiment above. We also experimented using a Swedish AM for Kyrgyz and found that it performed significantly worse with a 51% WER, presumably because the phoneme correspondences and phonologies are more dissimilar. Table 2 compares phonemes between the three languages.

Our baseline Filipino-only system performs quite well on our Cebuano test set, but our new approach still beats it by 10.4% absolute WER. The Corsican model significantly outperforms the baseline (which typically recognizes related Italian words, such as "novembre" instead of Corsican "nuvembre") but still gets a relatively high WER of around 34.8%. The new Maithili model only slightly outperforms the baseline of 79.3%.

We believe that the higher WER in Maithili and Corsican are related to the relatively smaller sizes of their LMs. Inspecting the output for incorrectly recognized utterances, it appears that the AM is doing well, but there are out-of-vocabulary issues in the LMs which make it impossible for the decoding graph to emit the target words.[1] For example, the Maithili model misrecognizes "सर्यक" (suryak) as "सूर्य" (surya) and "मने" (maney) as the English word "many". "सर्यक" and "मने" do not exist in our vocabulary. In these cases, as elsewhere, the pronunciation of the recognized words is quite close to the truth, suggesting that the AM is holding up reasonably well.

To see if adding more text data would help, we added in some more Maithili text data from a small corpus that we acquired externally. With a vocabulary size of 11K words, our Maithili WER was originally at 81.6% (worse, in fact, than the 79.3% baseline). After adding in some more text data to bring our model to a total vocabulary of 19K words, the WER decreased to 74.8%. We then injected more unigram data, bringing the LM size to about 35K words, further lowering the WER to 73.2%, which is the best result we were able to achieve for Maithili. We explored using a much larger Maithili wordlist from FastText [29] as well, but this causes a slight regression to 73.8%, which we believe is because of the high noise levels in this data source. In the end, we were able to bring our Maithili WER down by about 8.4% absolute by just adding a small quantity of additional high-quality training data for the LM. In improving WER, sentence data from which we could derive n-grams was more helpful than unigram data.

Our results suggest that our approach works well for languages such as Kyrgyz and Cebuano with reasonably large text corpora. In languages with less text data, WER is higher, but significant accuracy gains can be obtained by injecting more text data to increase the vocabulary of the models, and to get better n-gram probability estimates. Fortunately, text data is significantly easier to obtain than audio data.

[1] We also noticed some transcription quality issues with the golden transcripts in our Maithili test set, which may artificially inflate the WER somewhat.

3.2.2. Comparing with adaptation approaches

Above, we showed that simply evaluating our Marathi and Gujarati test sets on our baseline Hindi set-up yielded a significantly higher WER than using a tailored set-up. Of course, it could be equally worthwhile to look into adapting this Hindi set-up in different ways, e.g. by adapting the AM or the LM towards the target language. We ruled out doing any adaptation for the AM, since it would require collecting at least some audio training data in the target language. For the LM, one approach could be to use our existing text corpus to apply a big biasing model, using the approaches described in [30] to the LM in the source-language recognizer. For example, we could apply a Cebuano biasing model to the LM in our Filipino recognizer instead. However, this does not seem to work: we tried using 5K, 50K and 200K n-grams for the Cebuano biasing model on top of our Filipino LM, and obtained WERs of 27.7%, 58.8% and 128.3% respectively. Future work could explore this further, but it is worth noting that this approach cannot easily account for differences in grapheme-to-phoneme correspondences.

4. Future work

Based on our results, we believe that re-using existing AMs wholesale to build ASR systems in phonologically similar languages holds promise for bringing ASR systems to more languages. In terms of future work, we see three main areas to explore further: improved LMs; automatic candidate identification; and how to apply this approach in neural sequence-to-sequence ASR.

4.1. Language modeling improvements

In the target languages we evaluated, it appears that large gains could be achieved by making further improvements to the LMs. One clear way to do so would be to increase the size of our text corpora, but it may also prove fruitful to explore other modeling paradigms such as neural LMs. To create bigger text corpora, better mining approaches could be developed. Another option may be applying optical character recognition to digitize existing books or newspapers. Even including additional words from lexicographic dictionaries could yield further data, although one problem would be that dictionaries tend to include only headwords (e.g. "eat" but not "eaten"). This means that most words that are formed through morphological processes would not be covered by such a data source - even if they appear in natural speech. This morphological expansion problem may be alleviated as more languages are added to UniMorph [31].

4.2. Automatic candidate identification

Another area of future work would be to create a system that could automatically identify promising target languages, based on a list of source languages with existing AMs. Our linguists used only simple heuristics such as genetic and areal connections between languages, and basic information on the target-language phonemic and orthographic systems. A first attempt at building an automatic system could simply rank language pairs based on such basic features from [32].

However, a robust ranking system integrating phonemic and graphemic information seems much harder to build. With libraries such as FonBund [26] providing easy access to phoneme-inventory databases such as Phoible [25], it should be possible to automatically analyze the phoneme inventories of the world's languages to find further candidate pairs. But

Table 1: *Experiments in Four Languages*

Lang	Language Family	Speakers	Script	AM Source	LM Vocabulary	Baseline	WER
Cebuano	Austronesian	40M	Latin	Filipino	1M	28.5	**18.1**
Kyrgyz	Turkic	4.3M	Cyrillic	Turkish	1.1M	109.0	**18.7**
Corsican	Indo-European (Romance)	150K	Latin	Italian	26K	102.4	**34.8**
Maithili	Indo-European (Indic)	35M	Devanagari	Hindi	35K	79.3	**73.2**

Table 2: *Phoneme mappings between Kyrgyz, Turkish and Swedish*

Kyrgyz	ɑ	ɑː	b	d	dʒ	e	f	g	ɢ	i	iː	j	k	l	m	ɯ	ɯː	n	ŋ	ø	øː	o	oː	p	q	r	s	ʃ	ʃː	t	ts	tʃ	u	uː	v	χ	y	yː	z
Turkish	ɑ	**a**	b	d	dʒ	e	f	g	**ɟ**	i	iː	j	k	l	m	ɯ	ɯː	n	**n**	ø	øː	o	oː	p	**k**	**ɾ**	s	ʃ	**ʃ**	t	**tʃ**	tʃ	u	uː	v	**h**	y	yː	z
Swedish	**a**	ɑː	b	d	**d**	**eː**	f	g	**h**	**ɪ**	iː	**j**	k	l	m	**ʉː**	**ʉː**	n	ŋ	ø	øː	**ɔ**	oː	p	**k**	r	s	**ʂ**	**ʂ**	t	**ʈ**	**ʈ**	**ʊ**	uː	v	**ɧ**	**ʏ**	yː	**s**

this would not be as straightforward as just comparing the phoneme inventories to identify cases with large amounts of overlap. For example, if two languages both have 30 phonemes, with 28 phonemes perfectly overlapping but 2 widely different phonemes, this might be more problematic than having a pair where out of 30 phonemes, only 25 phonemes are shared. This may be counter-intuitive, but it would be easy to imagine a situation in which that these 5 phonemes differ only slightly, say in only one phonetic feature (such as aspiration). Our hypothesis is that this would be less of an obstacle to using a preexisting AM than having a few widely differing phonemes, so any difference metric here would have to factor in these phonetic features.

The usage frequency for every individual phoneme also plays a role: if some phonemes included in the target-language phoneme inventory are marginal and appear rarely, having a source AM that is not strong at distinguishing such phonemes may not be a problem. However, as far as we are aware, there are no large-scale databases that include within-language phoneme frequency. Theoretically, one could be created by taking text corpora across many languages and turning them into phonemic transcriptions to compute phoneme frequency statistics, but this would be challenging due to the unavailability of reliable grapheme-to-phoneme (G2P) conversion models. Another problem would be allophonic variants, as in Hawaiian where the voiceless alveolar stop [t] and the voiceless velar stop [k] are two ways of realizing a single phoneme. Information on allophones would be needed to estimate how well an AM would fit, but it is again not typically available in large-scale databases.

It is possible to run more experiments by manually identifying candidate pairs. But scaling such analyses across the world's languages will require a more systematic approach. Automatic methods may also help identify languages that are highly dissimilar from languages with existing audio corpora, and therefore would be especially worth creating large audio corpora for. These corpora could then form the foundation for ASR systems for similar languages through simple AM re-use.

4.3. End-to-end modeling

Recently, end-to-end seq2seq approaches to build ASR systems, such as recurrent neural network transducers (RNNTs), has become commonplace. RNNTs consist of an encoder and a decoder, where the roles of the encoder and decoder are roughly equivalent to that of an AM and LM in a traditional system. The encoder and the decoder are jointly trained on a training set of transcribed audio utterances. While promising results can be achieved in low-resource languages by training multilingual RNNTs [33], possibly combined with techniques such as LM fusion [34], no work appears to have been published on using RNNTs for languages with no audio training data.

This is perhaps not very surprising, since seq2seq models typically produce output units at grapheme, word-piece, or word level. While fusion with a target-language LM would be possible, it seems non-trivial to make our approach work in the seq2seq paradigm, unless the target-language G2P relationship is basically identical to the source-language G2P correspondences, and unless the target language uses only exactly the same graphemes or a subset thereof as the target language. Imagine a hypothetical case where we take an RNNT built for Spanish, which uses the Roman alphabet. If we re-spell all the transcripts in the test set into the Cyrillic alphabet, using 1:1 mappings, it seems challenging to make the RNNT produce the correct output, unless we show the RNNT some training examples of Spanish audio with Cyrillic text; but this requires audio in the target language, and undermines our goal.

5. Conclusion

We have shown that it is possible to build large-vocabulary ASR systems for languages with no audio training data at all, as long as a high-resource language that is sufficiently phonologically similar can be identified. Simply re-using existing AMs from a phonologically similar source language without any modification, coupled with target-language pronunciation and language models, can achieve good results. Our method can achieve WERs below 20% at low cost: we only need to gather a handful of components to build an ASR system in a new language, namely an LM, a pronunciation model, some basic verbalizers, and optionally a small audio test set. These components are much easier to obtain than a large audio training set. Our approach is limited to languages where a sufficiently similar language already has an ASR system. However, in our preliminary assessment, this could be the case for many languages that currently do not have ASR systems available. We also identified areas of future work that we believe will help us increase the scalability of this approach further, as well as ways in which we can further improve the experimental research systems we built.

6. Acknowledgements

We'd like to thank Haruko Ishikawa for leading the data collection efforts for our test sets; our linguistic team that created the pronunciation models; and Elnaz Sarbar, Ben Haynor, and Steven Knauer. We'd also like to thank Françoise Beaufays and Pedro Moreno for providing inspiration and executive support.

7. References

[1] Simons, Gary F. and Charles D. Fennig (eds.), "Ethnologue: Languages of the World," *Ethnologue: Languages of the World*, vol. 21, no. 3, 2018.

[2] M. J. F. Gales, K. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *SLTU*, 2014.

[3] P. Biggs, "The state of broadband: Broadband catalyzing sustainable development," International Telecommunications Union and Broadband Commission for Sustainable Development and UNESCO, Tech. Rep., 2017.

[4] C. van Heerden, M. Davel, and E. Barnard, "Medium-vocabulary speech recognition for under-resourced languages," 05 2012, pp. 146–151.

[5] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran, "Speech recognition for illiterate access to information and technology," 06 2006, pp. 83 – 92.

[6] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. A. Landay, "Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 159:1–159:23, Jan. 2018. [Online]. Available: http://doi.acm.org/10.1145/3161187

[7] A. Wills, G. Barrie, and J. Kendall, "Conversational interfaces and the long tail of languages in developing countries." [Online]. Available: http://dfslab.net/nlp-language-divide-html/

[8] M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–311, Jun. 1997. [Online]. Available: http://dl.acm.org/citation.cfm?id=972695.972698

[9] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Interspeech*, 2010.

[10] S. Bird, F. R. Hanke, O. Adams, and H. Lee, "Aikuma: A mobile app for collaborative language documentation," in *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 1–5. [Online]. Available: http://www.aclweb.org/anthology/W14-2201

[11] Common voice. Mozilla. [Online]. Available: http://voice.mozilla.org/en

[12] M. Prasad, T. Breiner, and D. van Esch, "Mining training data for language modeling across the world's languages," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 61–65. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-13

[13] 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.

[14] A. Deri and K. Knight, "Grapheme-to-Phoneme Models for (Almost) Any Language," in *Proc. ACL 2016: 54th Annual Meeting of the Association for Computational Linguistics*, Germany, August 2016, pp. 399–408.

[15] B. Peters, J. Dehdari, and J. van Genabith, "Massively Multilingual Neural Grapheme-to-Phoneme Conversion," in *Proc. of the First Workshop on Building Linguistically Generalizable NLP Systems*, Denmark, 2017, pp. 19–26.

[16] H. Sak, F. Beaufays, K. Nakajima, and C. Allauzen, "Language model verbalization for automatic speech recognition." in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8262–8266.

[17] K. Sodimana, P. D. Silva, R. Sproat, T. Wattanavekin, A. Gutkin, and K. Pipatsrisawat, "Text normalization for Bangla, Khmer, Nepali, Javanese, Sinhala and Sundanese text-to-speech systems," in *SLTU*, 2018, pp. 147–151.

[18] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," *CoRR*, vol. abs/1511.06066, 2015. [Online]. Available: http://arxiv.org/abs/1511.06066

[19] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(00)00094-7

[20] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," 03 2010, pp. 4334–4337.

[21] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system," in *INTERSPEECH*, 2009.

[22] N. T. Vu, F. Kraus, and T. Schultz, "Rapid building of an asr system for under-resourced languages based on multilingual unsupervised training," in *INTERSPEECH*, 2011.

[23] X. Li, S. Dalmia, D. R. Mortensen, F. Metze, and A. W. Black, "Zero-shot learning for speech recognition with universal phonetic model," 2019. [Online]. Available: https://openreview.net/forum?id=BkfhZnC9t7

[24] C. Gooskens, V. Van Heuven, J. Golubovic, A. Schüppert, F. Swarte, and S. Voigt, "Mutual intelligibility between closely related languages in europe," *International Journal of Multilingualism*, pp. 1–25, 07 2017.

[25] S. Moran, D. McCloy, and R. Wright, *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014. [Online]. Available: http://phoible.org/

[26] A. Gutkin, M. Jansche, and T. Merkulova, "FonBund: A Library for Combining Cross-lingual Phonological Segment Data," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[27] A. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath, and K. Rao, "Acoustic modelling with CD-CTC-SMBR LSTM RNNS," in *ASRU*, 2015.

[28] M. Chua, D. V. Esch, N. Coccaro, E. Cho, S. Bhandari, and L. Jia, "Text Normalization Infrastructure that Scales to Hundreds of Language Varieties," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018.

[29] E. Grave. (2017, October) Language identification. Accessed: June 4, 2018. [Online]. Available: https://fasttext.cc/blog/2017/10/02/blog-post.html

[30] P. S. Aleksic, M. Ghodsi, A. H. Michaely, C. Allauzen, K. B. Hall, B. Roark, D. Rybach, and P. J. Moreno, "Bringing contextual information to google speech recognition," in *INTERSPEECH*, 2015.

[31] J. Sylak-Glassman, C. Kirov, D. Yarowsky, and R. Que, "A language-independent feature schema for inflectional morphology," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

[32] H. Hammarström, R. Forkel, and M. Haspelmath, "Glottolog 3.3," Max Planck Institute for the Science of Human History, Jena, 2018. [Online]. Available: https://glottolog.org/ accessed 2019-03-18

[33] S. Toshniwal, T. N. Sainath, R. Weiss, B. Li, P. Moreno, E. Weinsten, and K. Rao, "Multilingual speech recognition with a single end-to-end model," 2018. [Online]. Available: https://arxiv.org/pdf/1711.01694

[34] 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," *CoRR*.