



Cross-dialect lexicon optimisation for an endangered language ASR system: the case of Irish

Liam Lonergan¹, Mengjie Qian²,
Neasa Ní Chiaráin¹, Christer Gobl¹, Ailbhe Ní Chasaide¹

¹Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland

²Engineering Department, Cambridge University, UK

mq227@cam.ac.uk, {berthelh, llonerga, murpha61, nichiar, cegobl, anichsid}@tcd.ie

Abstract

Lexicon optimisation strategies, addressing the problem of dialect divergence, are tested in an ASR system for Irish. As in many endangered languages, Irish has no spoken standard, but rather, three very different dialects of Ulster (Ul), Connaught (Co) and Munster (Mu). Furthermore, the complex sound system and ancient, opaque writing system result in sound-to-grapheme mappings that differ considerably across dialects. A hybrid ASR system was trained on (predominantly) native speaker speech data, balanced across the dialects. Experiment 1 tested whether a Global lexicon, which captures dialect variant forms with relatively abstract representations, can perform as well as a Multi-dialect lexicon containing all dialect variants. Three dialect-specific lexicons were also included in the tests. The Global lexicon did yield the best performance and experiment 2 tested whether further reductions to its phoneset might further enhance its performance. These included (i) merging a Tense-Lax contrast among coronal sonorants, not common to all dialects, and (ii) merging the contrast of voiceless-voiced sonorants, as the voiceless member is relatively infrequent. Results showed but a slight enhancement and only for Mu dialect, which is the one most aligned to the phoneset reduction.

Index Terms: Irish, speech recognition, cross-dialect variation, lexicon, minority language

1. Introduction

One of the many challenges in developing speech technologies for endangered languages stems from the fact that there may be multiple diverging dialects, but no spoken standard variety. This is the case for Irish: there is no spoken standard and 3 main dialects which differ considerably in pronunciation, the dialects of Ulster (Ul), Connaught (Co) and Munster (Mu). As a way of dealing with this challenge, different approaches to lexicon building are tested by this study in a recently developed hybrid ASR system for Irish. A strategy adopted here was to build a Global lexicon, i.e., a compact lexicon where relatively abstract representations are used to capture dialect-varying forms. In Experiment 1, we hypothesised that this Global lexicon would perform as well as a Multi-dialect lexicon, a more standard approach which simply includes the phonemic representations for all dialect forms. It was of course expected that both of these lexicons would provide much better results overall than lexicons which reflect one or other of the three dialects.

The best performance was indeed obtained with the Global lexicon, and in Experiment 2, two further modifications to this lexicon were implemented. These involved reducing the phoneset of the Global lexicon by merging contrasts that are either

not common to the three dialects or contrasts that are common across the dialects, but which are relatively infrequent in their occurrence. The sociolinguistic context and its implications for Irish speech technology development are explained in the following section, while Section 3 provides the motivation for the present experiments, explaining the challenges of cross-dialect variation and how it interacts with the complex sound system and the opaque mapping of sound to orthography in Irish.

2. Sociolinguistic Background

The present study is part of a broader initiative, ABAIR [1], which aims to develop linguistic resources, speech technologies and applications deploying both technologies and resources for Irish. Unlike mainstream technology development for the major languages, speech technology development for Irish is not driven by commercial considerations, but rather must be sensitive to the needs of the language community. This sociolinguistic backdrop to the present approach is outlined briefly here.

Irish is a Celtic language, which, although endangered [2], is the first official language of the State – something unusual for a minority language. It is spoken as a community language in Gaeltacht areas, scattered mainly on the western seaboard. There is no spoken standard: the evolution of a spoken (and written) standard was truncated with the overthrow of the Gaelic system in the early 17th century. The subsequent decline of the language, accelerated by a catastrophic famine in the mid 19th century and the mass emigration that followed, left shrinking Irish speaking communities in increasingly isolated linguistic ‘islands’. These dialects have evolved differently, in vocabulary, syntax and most strikingly, in pronunciation. Today there are 3 main dialects, referred to here as the dialects of Ulster (Ul), Connaught (Co) and Munster (Mu). As Ireland’s first official language, Irish is taught to all at primary and secondary school levels, so that there is a population of non-native (L2) speakers, with varying levels of proficiency.

This sociolinguistic context is important for technology building. For the major world languages, the building of TTS and ASR systems would have focused initially on the standard variety, with subsequent efforts directed at extending coverage to non-standard varieties. However, for Irish, as for minority and endangered languages which have no spoken standard, a prerequisite is to cater from the outset for the multiple native dialects. As speech and language technology can play a vital role in documenting, maintaining and revitalising endangered languages [3], the most pressing need is for applications that support these endangered language communities. Thus, in building text-to-speech synthesis systems, multi-dialect provi-

sion was the earliest priority¹. Equally, a priority for the initial ASR system is that it should cater for the divergent pronunciation of the three main dialects and allow applications that are directed at these communities.

With these considerations in mind, the ASR system presented here was built with speech corpora recorded predominantly by native speakers (L1) and balanced across the three dialects. A relatively small proportion of L2 speech data was included (see Section 5.1.1). In the longer term, it will be important to cater also for L2 language speakers, particularly as there will be many applications for language learners, within the schooling system and beyond. Although the present ASR system focusses on maximising its potential use for the native speaking communities, it was nonetheless of interest to see how well it might perform with L2 speech. As the Test set used in the present study draws on two sources of data, featuring both native speakers and L2 speakers, it allows some inferences to be drawn on this issue.

3. Cross-dialect Variation

To explain the complications of cross-dialect variation, we note some basic features of the sound system and the sound-to-orthography mapping of Irish.

The sound system of Irish is complex. Its most fundamental feature is a consonant-quality contrast of velarised (C^v) vs. palatalised sounds (C^j), as illustrated in Figure 1 (for Ulster Irish [4], [5] and [6]). Further somewhat unusual features are the opposition of Tense and Lax coronal laterals and nasals (Lax members shown in blue font in Figure 1). This opposition is not identical across the three dialects: if we illustrate in terms of the laterals (and for now ignore the voiceless laterals), in the Ulster dialect there may be a 4-way opposition, although the fourth member (shown in brackets) is marginal (see [5] and [6]); in Connaught there is a 3-way opposition; in Munster the system is further reduced to a 2-way opposition. There is also a contrast between voiced and voiceless sonorants (laterals, nasals and rhotics). Although this contrast is common to all the dialects, the voiceless members (red font in Figure 1) are fairly infrequent. The simplification of the Lexicon tested in Experiment 2 below entails (i) merging of the Tense-Lax contrast of coronal nasals and laterals and (ii) merging of the voiceless-voiced contrast of all sonorants.

	Labial	Dental	Alveolar	Alv-Pal	Palatal	Velar	Glottal
Stop	p ^v b ^v p ^j b ^j	t ^v d ^v		t ^h d ^h	c ɟ	k g	
Fric/App	f ^v w ^v f ^j v ^j		s ^v	ç	ç j	x ɣ	h
Nasal	ɲ ^v m ^v ɲ ^j m ^j	ɲ ^v ɲ ^v	ɲ ^v n ^v (ɲ ^j n ^j)	ɲ ^h n ^h ɲ ^h n ^h		ŋ ɳ	
Lateral		l ^v l ^v	(l ^v l ^v) l ^j l ^j	l ^h l ^h l ^h l ^h			
Rhotic			r ^v r ^v r ^j r ^j				

Figure 1: *Consonants of Irish, UL dialect (based on [4], [5] and [6]). The phonemes reductions in Experiment 2 indicated in blue (Lax nasals and laterals) and red (voiceless sonorants) font.*

The orthographic system of Irish uses the Latin alphabet. It provides a poor match to the rich system of consonantal

¹TTS systems for the 3 main dialects of Irish are available at www.abair.ie

contrasts, and the quality contrast of the consonants is not directly notated in the orthography. Rather, consonantal quality is indirectly inferred from vowel letter(s) adjacent to the consonant: an adjacent ‘i’ or ‘e’ letter borders a palatalised consonant, whereas adjacent ‘a, o, u’ letters serve to denote that the adjacent consonant is velarised. For example, in the word *buíon* (band), the vowel letters ‘u’ and ‘o’ indicate that the initial and final consonants are velarised, the letter ‘í’ denotes the actual vowel quality ([i:]) of the syllable nucleus. This orthographic convention results in what can look like bewildering strings of vowel letters where there is a single vowel phoneme: depending on the consonantal context, written sequences such as ‘áí, oí, aoi, uí, oí, ígh, ao’ all map to [i:] in Ulster Irish. The opacity of written Irish arises also from its antiquity. The written language maintains forms whose pronunciations have changed over many centuries, evolving differently in the different dialects. The cross-dialect difference in pronunciation may be at the level of an individual phone (the grapheme ‘ao’ is realised as either [i:] or [e:] depending on dialect), and may also involve grammatical morphemes such as the verbal future morpheme affix, written as -faidh or -fidh, and which, depending on the dialect, may be pronounced as [hi], [hə] or [hiʃ] with little correspondence to the spelling in any instance.

4. Global and Multi-dialect Lexicons

Experiment 1 compares five possible lexicons in the Irish hybrid ASR system, a Global lexicon (Glb), three dialect-specific lexicons, and a Multi-dialect lexicon (Multi) which includes all variants present in the latter three. Of key interest here is the performance of the Global lexicon, relative to the Multi-dialect lexicon.

Table 1: *Entries for the Ulster, Connaught, Munster, Global and Multi-dialect lexicons. Global items include abstract units.*

Orth	Multi	Co	Mu	Ul	Glb
saol	s ^v i:l ^v s ^v e:l ^v	s ^v i:l ^v	s ^v e:l ^v	s ^v i:l ^v	s ^v AOI ^v
adharc	e:rk ^v airk ^v	airk ^v	airk ^v	e:rk ^v	ADHrk ^v
amhrán	o:r ^v an ^v aur ^v an ^v	o:r ^v an ^v	aur ^v an ^v	o:r ^v an ^v	AMHr ^v an ^v
saolaíodh	s ^v i:l ^v iu s ^v i:l ^v iəx s ^v e:l ^v iəx	s ^v i:l ^v iəx	s ^v e:l ^v iəx	s ^v i:l ^v iu	s ^v AOI ^v iODH#
dúinfaidh	d ^v u:n ^v hi d ^v u:n ^v hə d ^v u:n ^v hɪʃ	d ^v u:n ^v hə	d ^v u:n ^v hɪʃ	d ^v u:n ^v hi	d ^v un ^v IDH#

The Global lexicon, illustrated in Table 1, is inspired by earlier concepts from dialectology and it is in some respects akin to a proposal for a ‘common core’ description of Irish phonology [7]. Its initial formulation differs however from the latter in being maximally inclusive, e.g., by including phonemic contrasts that might be found in one dialect, but not in another (such as the Tense-Lax contrast of sonorants). It was an approach that we initially used to deal with cross-dialect variation for Irish speech synthesis: details of its formulation can be found in [8]. The Global lexicon contains all variants involving letter-to-sound mappings that are common to the dialects. Where dialect pronunciations diverge it provides abstract units, shown in CAPS in Table 1. These abstract units can represent a phoneme or longer strings of phonemes. For example, the Global unit AO corresponds to either the phoneme /i:/ or /e:/ depending on

the dialect. Abstract Global units can also represent grammatical morphemes (involving longer strings of phonemes) whose pronunciations diverge, such as the word final morphemes that mark verbal tense, mood and person. Examples of the latter in Table 1 are: ODH# (a verbal past tense marker, realised as [u] or [əx]); and IDH#, a verbal future tense marker which can yield either [hi], [hə] or [hrj] (mentioned in Section 3), depending on dialect. To sum up, the Global lexicon offers a compact, trans-dialect form of representation, allowing for rather abstract representation of those graphemes (phonemes or morphemes) whose pronunciations differ. Note that the abstract units of the Global lexicon are akin to (one of) the orthographic representations, and as such, represents a kind of half way house between phonemic and orthographic transcriptions.

Table 2: Number of phones/abstract units (#phn) and entries (#lex) in Dialect-specific, Global and Multi-dialect lexicons.

	Co	Ul	Mu	Multi	Glb	Glb_mdf
#phn	65	85	63	117	92	74
#lex	550k	549k	550k	1006k	540k	540k

As illustrated in Table 2, the Global lexicon is considerably more compact than the Multi-dialect lexicon and should have advantages in the size of the decoding lattices and the efficiency with which they can be searched. As the end goal here are applications, the latency of the system will be an important factor in many cases. Provided it can compete with the performance of a Multi-dialect lexicon, the Global lexicon offers further advantages going forward, such as facilitating the extension to further dialects and facilitating the coverage of the extensive grammatical inflections of Irish.

5. Data

5.1. Speech Corpora

Details of the speech datasets used for the Train and Test sets are given in Table 3. Efforts were made to balance both the Train and Test sets according to the dialect of the speakers (see Table 4). This act of balancing also meant that speakers lacking dialect information were not included, and resulted in a subset of our complete recording sets being used in this paper.

Table 3: Details of the speech datasets used.

dataset	#wav	#spk	#words	#vocab	#dur (h)
Train	39,609	281	338,643	15,018	37.24
Test	1174	20	8224	2103	1.14

5.1.1. Training data

The dialect-balanced speech corpora used to train the acoustic models included firstly, the ABAIR synthesis recordings of 7 L1 speakers for the three dialects, totalling 18.4 hours. Materials were dialect appropriate, e.g., books composed in the local idiom, and were read, sentence by sentence. A further corpus of 256 speakers (17.96 hours) reading from dialect-appropriate stimuli was derived from internal live and crowd-sourced recordings, using the ABAIR MíleGlór website². These were predominantly L1 speakers, but included a small number of L2 speakers (see below). An additional 1.1 hours of spontaneous conversational speech data was also included. This corpus consisted of L1 speakers. The recordings

²<https://www.abair.tcd.ie/studio/ga/recorder/>

were cleaned manually to remove utterances with non-speech, e.g. noise, laughter, etc. A total of 37.2 hours of speech was used to train the ASR acoustic model. The breakdown for speakers/words/vocabulary is provided in Table 3.

70% of informants (over 85% of Train set duration) declared themselves as native speakers and the remaining as L2. The proportion of duration for different dialects was: Ul 34.6%; Co 33.6%; and Mu 31.8%.

Table 4: Breakdown of Train and Test sets by dialect in hours

Dialect	Train		Test	
	#spk	#dur (h)	#spk	#dur (h)
Mu	103	11.8	7	0.45
Co	110	12.5	6	0.27
Ul	68	12.9	7	0.42
Total	281	37.2	21	1.14

5.1.2. Test data

The Test set was taken from two platforms. Firstly, recordings from the Mozilla Common Voice crowd-sourcing platform were used, but only where dialect preferences were specified and where predominantly positive listeners’ judgements were obtained. Upon listening, the speakers of these recordings were all deemed to be L2. The resulting corpus was both small (appr. 31m dur) and imbalanced for dialect. An additional 40m of L1 speech, collected using the MíleGlór platform was added, allowing us to balance our test set for dialect. There was no overlap in transcriptions or speakers with the data used in the Train set. The dialect breakdown of the Test set is given in Table 4.

5.2. Text Corpora

Different text corpora were combined to build the language models. The Corpus of Irish for Lexicography [9] (version 2021.1) developed by Gaois, DCU, with funding from Foras na Gaeilge, is referred to as Text A (72m words, 1.5m vocabulary). A version of the National Corpus of Irish, provided by Foras na Gaeilge, is referred to as Text B (52m words, c.0.25m vocabulary). The text from a spontaneous speech corpus of Irish is used and is referred to as Text C (c.4m words, c.0.08m vocabulary). Finally, Irish language text collected from Wikipedia is referred to as Text D (c.2.5m words, 0.13m vocabulary).

6. Experimental Set-up

The same basic set-ups were applied to the systems in our experiments with different lexicons, and will be detailed here. All the experiments are done using the Kaldi toolkit [10].

The acoustic model (AM) in the baseline ASR system was a Time-Delay Neural Network (TDNN) [11, 12], trained on the 37.2 hours Train set (Table 3) for 10 epochs. The initial alignment was produced by a triphone GMM-HMM trained with standard MFCC features, applying linear discriminative analysis (LDA), maximum likelihood linear transformation (MLLT), feature space maximum likelihood linear regression (fMLLR) and speaker adaptive training (SAT). The features for training the TDNN model were 40-dimensional high-resolution MFCCs stacked with 100-dimensional online extracted i-vectors. Two widely used on-the-fly data augmentation techniques for ASR – speed perturbation [13] and spectral augmentation (SpecAug) [14], were applied to augment the AM training data. The data was tripled with speed perturbation, and the warping factors for the speed perturbation were

0.9, 1.0 and 1.1. SpecAug operates on the log mel spectrogram of an audio clip. By randomly masking bands on the frequency domain and time domain, this methods leads to impressive improvements [14]. The TDNN model consists of 13 factorized TDNN (TDNN-F) layers with a size of 1024 and a bottleneck size of 128. It was trained with lattice-free maximum mutual information (LF-MMI) [15].

The baseline language model was a 3-gram model [16] trained on a combination of TextA TextB TextC and TextD, using the SRILM toolkit [17]. It has been shown beneficial to use lattice-rescoring [18, 19] with recurrent neural network language models (RNNLM) [20, 21]. In the experiments, an RNNLM was trained on TextC and TextD, then it was used to rescore the hypotheses generated from the 3-gram LM.

System fusion is a common method to make use of multiple similar systems and achieve a stable performance. As the number of training epochs affects how much a system is finetuned to the training data and as such how robust it will be to unseen testing data, fusion of variants of acoustic models trained using a different number of epochs was explored in the experiments.

7. Results and Discussion

Table 5 presents the results for both experiments 1 and 2, comparing in a) the Overall WER results for: single systems, which are trained for 10 epochs using the baseline 3-gram LM; fused systems using the baseline 3-gram LM; and fused systems rescored using an RNNLM. In section b) of Table 5, we look at the breakdown of Overall WER for the RNNLM rescored fused systems, according to the dialect affiliation of the speakers. It should be noted that the dialect affiliation here refers to the actual dialect for native speakers (L1), but simply to the dialect preference of L2 speakers, whose speech may approximate to dialect norms in varying degrees (see more on this below). 5c) compares WER for the two different corpora used in the Test set, MíleGlór and Mozilla. This is of particular interest because the percentage of L1 speakers in the former is 80% and there are none in the latter.

7.1. Experiment 1: testing Global vs Multi-dialect lexicons

A word list was created from the text corpora used for language modelling and transcribed with the appropriate ruleset for the Dialect-specific lexicons (1 pronunciation per word). The Multi-dialect lexicon included all dialect-specific variants (1.97 pronunciations per word). The Global lexicon (1 pronunciation per word) was created from the Global ruleset and is of particular interest here. Each lexicon was used to train an ASR system following the training pipeline detailed in Section 6.

Even though the lexicons have different amounts of phonetic units, which corresponds to different numbers of parameters in the acoustic model to be trained, while the size of the acoustic training data remains constant, it is still a fair comparison because the models are trained using a range of epochs to find the best system, in addition to the randomness of SpecAug.

In Table 5a) for the best performing systems i.e. RNNLM rescored, the Global lexicon yields the lowest WER. The Multi lexicon WER is only slightly higher and they both outperform the dialect-specific lexicons. In Table 5b), looking at the two lexicons of particular interest here (the Multi and the Global lexicons), the Global lexicon yields a better performance for Co and Ul speakers, but the Multi performs better for Mu. Unsurprisingly, the lexicons for individual dialects yield the lowest WER for Co and Ul. Curiously, this is not the case for Mu,

Table 5: WER% for all lexicons in Experiments 1 and 2. a) Overall WER% for all test speakers; b) breakdown of WER according to dialect affiliation of speakers; and c) breakdown of WER for the two corpora used in the Test set.

Exp. 1						Exp. 2
a)	Co	Mu	Ul	Multi	Glb	Glb-M
Single	13.72	13.66	13.64	13.1	13.38	13.56
Fused	12.83	12.97	12.91	12.60	12.50	12.83
+RNNLM	9.23	9.31	9.25	8.85	8.78	8.74
b)	Co	Mu	Ul	Multi	Glb	Glb-M
Co spk	9.73	9.92	10.54	10.35	9.98	9.98
Mu spk	8.10	7.84	8.34	6.96	7.31	7.22
Ul spk	10.17	10.58	9.58	10.11	9.74	9.74
c)	Co	Mu	Ul	Multi	Glb	Glb-M
MíleGlór	8.19	6.87	6.87	6.38	6.73	6.61
Mozilla	10.00	11.12	11.01	10.68	10.30	10.32

but we note that WERs for this dialect affiliation are the lowest across the board. In Table 5c) where the MíleGlór and Mozilla corpora are compared, there is a consistently large WER difference, being consistently higher in the latter corpus. Given that the speakers in this case were L2 speakers and that the speakers in the MíleGlór corpus were predominately L1 speakers, this may point to a better performance for L1 speech. However, this interpretation is tentative, as the corpora differed in terms of the control over recording environments and the materials recorded.

7.2. Experiment 2: reducing the Global lexicon

Experiment 2 examined whether a modified Global lexicon, which merged Tense-Lax and Voiced-Voiceless sonorants might further enhance its performance. As can be seen in Table 5, the modified Global lexicon yields mostly similar and sometimes better WER compared to the Global lexicon. However, as is clear in 5b), the only performance gain pertains to the Mu-affiliated speakers. This is not surprising as the reduction of the Tense-Lax contrast is a feature of this dialect (see Section 3).

8. Conclusions and Future Work

These results suggest that the Global lexicon proposed here presents a useful strategy for dealing with cross-dialect variation, by allowing a much more compact lexicon that equals or betters the performance of a Multi-dialect approach. This would have advantages to the efficiency and latency of the system, which would impact on future applications. Furthermore, it would facilitate the extension of the system to further dialects and help deal with the many inflectional morphology of Irish, which tends to vary across dialects, as illustrated in Table 1.

We are currently extending our speech corpora, with a concerted effort to record in the local Gaeltacht communities. We will also focus on building L2 speech corpora, exploring further how our systems perform across the L1-L2 divide. Further to the current approach, we are also investigating the potential of End-to-End ASR systems [22, 23] for dealing with the large variation in Irish speech, including the use of pretrained models, such as Wav2Vec 2.0 [24].

9. Acknowledgments

This work is part of the AB AIR initiative, which is supported by *An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán*, with funding from the National Lottery, as part of the *Straitéis 20 Bliain don Ghaeilge*.

10. References

- [1] ABAIR. Webpage with Irish Language Text-to-Speech Systems for Three Dialects. [Online]. Available: <https://www.abair.tcd.ie/>
- [2] C. Moseley, *Atlas of the World's Languages in Danger*. Unesco, 2010. [Online]. Available: <http://www.unesco.org/languages-atlas/>
- [3] A. N. Chasaide, N. N. Chiaráin, H. Berthelsen, C. Wendler, A. Murphy, E. Barnes, and C. Gobl, "Leveraging Phonetic and Speech Research for Irish Language Revitalisation and Maintenance," in *ICPhS the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 2019.
- [4] A. Ní Chasaide, *Irish, in Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.
- [5] A. Ní Chasaide, "Laterals in Gaoth Dobhair Irish and Hiberno-English," in *Papers in Celtic Phonology*. N.U.U., Coleraine, 1979, pp. 54–78.
- [6] A. Ní Chasaide, "Acoustic Study of Laterals in Donegal Irish and Hiberno-English." Master's thesis, University College of North Wales, Bangor, 1979.
- [7] M. Ó Murchú, "Common core and underlying forms," *Ériu*, vol. 21, pp. 42–75, 1969.
- [8] B. Ó Raghallaigh, "Multi-Dialect Phonetisation for Irish Text-to-Speech Synthesis: a Modular Approach," Ph.D. dissertation, Trinity College Dublin, 2010.
- [9] M. J. Ó Meachair, B. Ó Raghallaigh, Ú. Bhreathnach, G. Ó Cleircín, and K. Scannell, "Corpus Creation for Lexicographical Research: Corpas Foclóireachta na Gaeilge," pp. 278–305, 2021.
- [10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi Speech Recognition Toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.
- [11] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3214–3218.
- [12] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3743–3747.
- [13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3586–3589.
- [14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 2613–2617.
- [15] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 2751–2755.
- [16] J. T. Goodman, "A bit of progress in language modeling," *Computer Speech and Language*, vol. 15, no. 4, pp. 403–434, 2001.
- [17] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SriIm at sixteen: Update and outlook," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 5, 2011.
- [18] X. Liu, X. Chen, Y. Wang, M. J. Gales, and P. C. Woodland, "Two efficient lattice rescoring methods using recurrent neural network language models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1438–1449, 2016.
- [19] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5929–5933.
- [20] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.
- [21] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. H. Cernocky, "RNNLM - Recurrent Neural Network Language Modeling Toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 196–201.
- [22] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7829–7833.
- [23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 5036–5040.
- [24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, pp. 12 449–12 460.