Speech Communication 136 (2022) 76–83

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Seamless equal accuracy ratio for inclusive CTC speech recognition ☆

Heting Gao [a,*], Xiaoxuan Wang [a], Sunghun Kang [b], Rusty Mina [b], Dias Issa [b], John Harvill [a], Leda Sari [a], Mark Hasegawa-Johnson [a], Chang D. Yoo [b]

[a] University of Illinois at Urbana-Champaign, IL, United States
[b] Korea Advanced Institute of Science and Technology, Daejeon, South Korea

ARTICLE INFO

Keywords:
Speech recognition
Fairness

ABSTRACT

Concerns have been raised regarding performance disparity in automatic speech recognition (ASR) systems as they provide unequal transcription accuracy for different user groups defined by different attributes that include gender, dialect, and race. In this paper, we propose "equal accuracy ratio", a novel inclusiveness measure for ASR systems that can be seamlessly integrated into the standard connectionist temporal classification (CTC) training pipeline of an end-to-end neural speech recognizer to increase the recognizer's inclusiveness. We also create a novel multi-dialect benchmark dataset to study the inclusiveness of ASR, by combining data from existing corpora in seven dialects of English (African American, General American, Latino English, British English, Indian English, Afrikaaner English, and Xhosa English). Experiments on this multi-dialect corpus show that using the equal accuracy ratio as a regularization term along with CTC loss, succeeds in lowering the accuracy gap between user groups and reduces the recognition error rate compared with a non-regularized baseline. Experiments on additional speech corpora that have different user groups also confirm our findings.

1. Introduction

End-to-end trainable deep neural networks have become the state-of-the-art architecture for automatic speech recognition (ASR). These networks can reduce word error rates to below 2%, even in an open-vocabulary task (Li et al., 2019; Zeyer et al., 2019) provided that the network is trained with a sufficiently large dataset. However, there are concerns that such neural networks do not exhibit equally good performance for different subgroups of the population.

ASR systems based on deep architectures trained on different English dialects report large disparity in word error rate between different dialects (Li et al., 2018; Winata et al., 2020). Similar disparities exist among Arabic dialects (Elfeky et al., 2018), between black vs. white American English speakers (Koenecke et al., 2020), and as a function of the age of the speaker (Picone, 1990).

A number of provisional attempts to improve performance of multi-dialect ASR systems have been made: enlarging training datasets for under-resourced dialects (Koenecke et al., 2020), conducting dialect-dependent training (Elfeky et al., 2018) and adaptation (Winata et al., 2020), and adding dialect related information as additional features (Li et al., 2018). These methods have addressed the error rate disparity as a provisional resource problem, rather than treating error rate variance as a problem equal in importance with the attainment of low average word error rate. Research in data mining has demonstrated that artificial intelligence (AI) trained in an unfair environment will learn the unfairness of its teachers unless specifically instructed not to do so (Kamiran and Calders, 2009; Calders et al., 2009; Hardt et al., 2016). Our study is inspired by methods in the AI fairness literature such as demographic parity, equal odds, and equal opportunity, that have been used to minimize discriminatory predictions based on race or gender. Inspired by those methods, we seek to design ASR that works well for all users: a goal that has been described as "inclusive speech technology" (Scharenborg, 2021). Our goal is motivated by two observations: (1) ASR may be a useful productivity tool (Désilets et al., 2008), (2) historic socioeconomic disparities correlate with dialect (Carter and Callesano, 2018). ASR that has low error rates for historically advantaged dialects and high error rates for historically disadvantaged dialects risks exacerbating existing socioeconomic disparities.

☆ This work was supported by the Korean Institute for Information & Communications Technology Planning & Evaluation (IITP), South Korea grant number 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data, South Korea. Opinions and findings are those of the authors, and are not endorsed by IITP.

* Corresponding author.
E-mail addresses: hgao17@illinois.edu (H. Gao), xw27@illinois.edu (X. Wang), sunghun.kang@kaist.ac.kr (S. Kang), rrmina@kaist.ac.kr (R. Mina), dias.issa@kaist.ac.kr (D. Issa), harvill2@illinois.edu (J. Harvill), lsari2@illinois.edu (L. Sari), jhasegaw@illinois.edu (M. Hasegawa-Johnson), cd_yoo@kaist.ac.kr (C.D. Yoo).

https://doi.org/10.1016/j.specom.2021.11.004
Received 31 December 2020; Received in revised form 28 July 2021; Accepted 9 November 2021
Available online 2 December 2021

In order to study the inclusivity of a speech recognition system, we collected English speech data from 7 public dialect corpora. We combined their data to form a multi-dialect corpus and treated each dialect as a user group. Experiments on our new corpus indicate that models trained with multiple dialects have unequal accuracy across dialects. Dialects with limited training data benefit when training on the joint training set, suggesting that small training datasets are one source of the accuracy gap. Among dialects with similar-sized training datasets, however, accuracy differences do not disappear even when we train separate models for each dialect, suggesting that some of the dialects we study are less homogeneous than others, and therefore harder to transcribe.

Our study attempts to reduce the performance disparity across different user groups. For this purpose, we propose a new measure derived from published measures of algorithmic fairness, which we call the equal accuracy ratio, and we integrate this measure into a standard CTC-based speech recognizer to reduce the performance disparity in ASR systems. The main contributions of this paper are:

- We create and publish a novel multi-dialect corpus by combining existing corpora in seven dialects of English (African American, General American, Latin American, British English, Indian English, Afrikaaner English, and Xhosa English).
- We propose "equal accuracy ratio", an inclusivity measure that is inspired by equal opportunity and can be seamlessly integrated into the loss function of a neural speech recognizer as a regularization loss to improve the inclusiveness of the ASR.
- We demonstrate that training with the equal accuracy ratio can improve both inclusiveness and accuracy of ASR with experiments conducted on the novel dialect corpus and two other speech corpora, namely CORAAL (Kendall and Farrington, 2018) and UASpeech (Kim et al., 2008), that have different user group partitions.

The rest of this paper is summarized as follows. Section 2 reviews the studies relevant to fairness and multi-dialect speech recognition. Section 3 introduces our proposed measure to improve the inclusiveness of ASR systems. Section 4 describes three corpora used for this study, including a new multi-dialect corpus created by the combination of seven different source corpora in seven racially, ethnically, and geographically diverse dialects of English. Section 5 describes experimental methods (neural network architectures and experimental setup). Section 6 presents experimental results. Sections 7 and 8 discuss and summarize our conclusions.

2. Related work

2.1. Fairness measures

A classifier trained on a corpus can be unfair toward certain groups of users due to historical bias or insufficient minority group training data. In one of the earliest studies of AI fairness, credit prediction models decided whether or not to accept a loan application (Kamiran and Calders, 2009). Trained models were age-discriminatory according to the criterion of demographic parity. The demographic parity criterion requires that there should be no difference between the average outcomes for different user groups:

$$|p_{\hat{Y}|A}(1|0) - p_{\hat{Y}|A}(1|1)| = 0, \tag{1}$$

where we define $p_{\hat{Y}|A}(y|a)$ to be the probability that the hypothesis result, $\hat{Y}$, takes value y, given that the protected attribute (e.g., age) has a value $A = a$. The demographic parity gap can be reduced by massaging dataset labels or giving more weight to samples from disadvantaged groups (Calders et al., 2009). A recent paper (Anahideh and Asudeh, 2020) proposes that demographic parity can select data to create a fair training set, instead of modifying the training labels in an existing dataset.

Demographic parity is less useful when there is a desired or ground truth result, Y, which is known, and which is correlated with the protected attribute. If the desired result is correlated with A, then imposing Eq. (1) reduces accuracy. When the ground truth is known and desirable, the equal odds and equal opportunity criteria (Hardt et al., 2016) are more desirable than demographic parity. The equal odds criterion requires conditional independence between hypothesis and attributes given ground truth, i.e.,

$$|p_{\hat{Y}|A,Y}(\hat{y}|0,y) - p_{\hat{Y}|A,Y}(\hat{y}|1,y)| = 0 \quad \forall y, \hat{y} \in \{0,1\}, \tag{2}$$

where Y is the ground truth and $\hat{Y}$ is the hypothesis. Equal odds requires that any particular mistake is made with equal probability, regardless of the setting of the protected attribute. "Equal opportunity" is a relaxation of equal odds, which focuses only on the error rate of the classifier: Eq. (2) is enforced only when the hypotheses match the ground truths ($y' = y$).

Predictive rate parity (Zafar et al., 2017) takes a different perspective, arguing the prediction should reflect the real performance of the group. The ground truth should be conditionally independent of group attributes given the predictions:

$$|p_{Y|A,\hat{Y}}(y|0,\hat{y}) - p_{Y|A,\hat{Y}}(y|1,\hat{y})| = 0 \quad \forall y, \hat{y} \in \{0,1\}. \tag{3}$$

These fairness requirements cannot all be simultaneously satisfied (Kleinberg et al., 2017); it is necessary for the users of a particular AI technology to decide which of these fairness criteria are the most desirable for their technology. Given such a fairness specification, a model can be trained to be fairer by optimizing the gap between predicted probabilities for each pair of groups. However, in their original published forms (as shown above), all of these criteria assume binary outcome variables ($\hat{Y}$). Extensions to real-valued and real-vector outcome variables have been published, but no previous study has published an extension to variable-length sequential outcome variables.

2.2. Speech recognition

End-to-end neural network based speech recognition systems can achieve very high performance given sufficient training data. State-of-the-art deep neural architectures for speech recognition combine acoustic and linguistic encoders with a mechanism for reducing the length of the sequence, from a larger number of input frames to a smaller number of output symbols (Schlüter, 2019). There is currently active research comparing the capabilities of CTC (Graves et al., 2006), attention-based encoder–decoder structures (Li et al., 2019), and hidden Markov models (Schlüter, 2019) for modifying the sequence length. CTC modifies the sequence length by ignoring repeated or blank phone symbols, thereby focusing the training procedure on a small number of conditionally independent frame classifications. Baidu's Deepspeech (Hannun et al., 2014) is a recurrent neural network (RNN)-CTC model that achieves high performance on both English and Mandarin. Recent successful CTC-based models include LipNet (Assael et al., 2016) and the DeepMind RNN-CTC model (Shillingford et al., 2018).

Denote the spectral features of the ith utterance as a set of frames $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}]$ where T is the number of frames. Denote the reference transcription as $y^{(i)} = [y_1^{(i)}, y_2^{(i)}, \ldots, y_{S_i}^{(i)}] \in \mathcal{Y}^+$, and the ASR output hypothesis as $\hat{y}^{(i)} = [\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_{\hat{S}_i}^{(i)}] \in \mathcal{Y}^+$, where S_i and $\hat{S}_i$ are the lengths of the reference and hypothesis transcriptions of ith sample and $\mathcal{Y}$ is the set of all transcription characters. The true conditional probability distribution $p_{Y|X}(y|x)$ is unknown; the ASR computes an estimated distribution $p_{\hat{Y}|X}(y|x)$ in order to minimize the cross-entropy of the training corpus,

$$\mathcal{L}_{CE} = -\sum_{i=1}^{|S|} \ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \tag{4}$$

where $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(|S|)}, y^{(|S|)})\}$ is a training corpus containing utterances with known transcriptions.

Connectionist temporal classification (CTC (Graves et al., 2006)) performs time-scale modification by positing an alignment sequence, $\Pi^{(i)} = [\Pi_1^{(i)}, \ldots, \Pi_T^{(i)}]$ whose instance value is $\pi^{(i)} = [\pi_1^{(i)}, \ldots, \pi_T^{(i)}]$. Each time-aligned character $\pi_t^{(i)}$ is either one of the transcription characters ($\pi_t = y_s$ for some s), or $\pi_t = \_$ where _ is a special "blank" character. For example, suppose we have a 5-character text "hello" ($S = 5$) encoded in a 14-frame speech waveform ($T = 14$); the transcription and alignment might be

$$y = [h, e, l, l, o], \pi = [h, h, e, e, \_, \_, l, l, l, \_, l, \_, o].$$

Training data are often provided with only the transcriptions, and the alignment information is not given. If the alignments are known, it would be easier to estimate the cross-entropy given in Eq. (4) by taking the sum of the log probabilities of the correct alignment at each frame.

Since alignment is not known, CTC computes the cross-entropy by marginalizing over all the possible alignments that can be mapped to the true transcription using a surjective time-compression function defined as:

$$\mathcal{B} : (\mathcal{Y} \cup \{\_\})^+ \rightarrow \mathcal{Y}^+.$$

A commonly used $\mathcal{B}$ first removes repeated labels and then removes all "blank" characters. For any valid alignment π, $\mathcal{B}(\pi)$ is a unique y. For any valid y, $\mathcal{B}^{-1}(y)$ is the set $\{\pi : \mathcal{B}(\pi) = y\}$. The log-probability of a transcription $y^{(i)}$ given the input frames $x^{(i)}$ can therefore be computed as

$$\begin{aligned}\mathcal{L}_{CTC}^{(i)} &= -\ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}),\\ &= -\ln \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} \exp(e_t(\pi_t)),\\ &= -\operatorname{logsumexp}_{\pi \in \mathcal{B}^{-1}(y)} \sum_{t=1}^{T} e_t(\pi_t).\end{aligned}$$

where $e_t(\pi_t)$ is the log output of a softmax layer predicting the transcription label at time t. The input of this softmax layer can be a bidirectional LSTM, Transformer, or other neural network parameterized by θ and having access to the whole sequence x.

The training loss is therefore the summation over CTC loss of each speech sample as

$$\mathcal{L}_{CTC} = \sum_{i=1}^{|S|} \mathcal{L}_{CTC}^{(i)}.$$

One concern with modern deep neural ASR models is that the model does not have equal performances for different user groups, potentially creating disparity of opportunity over region, age, gender, race, educational status, disability, class, etc. Experiments on the Japanese Newspaper Article Sentences corpus (Itou et al., 1999) show 10% higher word error rate for older voices than for younger voice (Vipperla, 2011). A study examining Youtube's automatic captions reports lower accuracy for female speakers (Tatman, 2017). Experiments on a neural ASR system trained using 7 different dialects of English from America, India, Britain, South Africa, Australia, Nigeria & Ghana and Kenya report large disparities in word error rate ranging from 10.6% for American English to 33.4% for Ghana & Kenya English in dialect-dependent training (Li et al., 2018). Recognition systems trained on different Arabic dialects (Egyptian, Gulf, Levantine and Maghrebi) suffer similar error rate disparities (Elfeky et al., 2018), ranging from 26.3% for Maghrebi Arabic to 34.0% for Egyptian Arabic. Recently, a study on state-of-the-art ASR systems from Amazon, Apple, Google, IBM, and Microsoft reports that all these systems have obvious racial disparities (Koenecke et al., 2020). The average word error rate for the black speakers is twice as large as that of the white speakers.

Methods proposed in improving these models include adding group-specific features in the training (Li et al., 2018), fine-tuning the model on data from each group of users, switching models based on group information (Yang et al., 2018), etc. These methods improve the accuracy for each user group and thus the overall accuracy of the model. However, they do not emphasize the inclusiveness of the ASR model. Our proposed method, the equal accuracy ratio, estimates the inclusiveness of the ASR during the training of a CTC-based sequence-to-sequence transcription model, and explicitly balances the relative importance of inclusiveness against average error rate over a given set of training corpora.

3. Equal accuracy ratio

The proposed equal accuracy ratio is an adaptation, to sequence-learning models, of the equal opportunity training criterion. Equal opportunity was defined in Hardt et al. (2016) as:

$$|p_{\hat{Y}|A,Y}(y|0, y) - p_{\hat{Y}|A,Y}(y|1, y)| = 0 \quad \forall y \in \mathcal{Y}. \tag{5}$$

There are 3 candidate definitions we can use for the purpose of adapting the equal opportunity criterion to ASR:

1. Matched frames: $p_{\hat{Y}|A,Y}(y|a, y)$ could be measured using sets of frames, with different values of the protected attribute $A = a$, for which the recognizer should output character y. However, matched frames would need a ground truth alignment, which are not required for CTC training, and are rare in practice.
2. Matched transcription: $p_{\hat{Y}|A,Y}(y|a, y)$ could be measured using sets of waveforms, with different values of the protected attribute, that have exactly the same transcription. Corpora that provide identical texts spoken by members of different groups exist (e.g., UASPEECH (Kim et al., 2008) and TIMIT (Lamel et al., 1986)), but are rare and small.
3. Matched accuracy: The sentence accuracy of an ASR, for user group a, is given by

$$p_{\hat{Y}|A}(Y|a) = \sum_{y} p_{Y|A}(y|a) p_{\hat{Y}|A,Y}(y|a, y). \tag{6}$$

Inclusiveness of an ASR might be reasonably defined to mean that accuracy is the same for different demographic groups, even if they do not say exactly the same things. The equal-accuracy marginalization of Eq. (5) is

$$|p_{\hat{Y}|A}(Y|0) - p_{\hat{Y}|A}(Y|1)| = 0. \tag{7}$$

We will use definition since it codifies the criterion that matters most to users (the accuracy of the speech recognizer), without requiring any additional constraints.

Extending equal opportunity in Eq. (7) to the multi-group case, the ASR provides equal opportunity if and only if

$$|p_{\hat{Y}|A}(Y|a) - p_{\hat{Y}|A}(Y|a')| = 0 \quad \forall a, a', \tag{8}$$

$$|\ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a')| = 0 \quad \forall a, a'. \tag{9}$$

Taking the logarithm on both sides of Eq. (9) does not alter the equality, but provides computational benefits as we will show shortly. After the manipulation, minimizing the gap on the left-hand side is actually forcing the ratio of accuracies to be one. Therefore we call the objective "**equal accuracy ratio**".

In practice, equal accuracy is rarely achieved. An ASR can be explicitly trained to minimize violations of equal accuracy, however, by training it to minimize the equal accuracy ratio, $\mathcal{L}_{EAR}$, defined as the total absolute difference between the cross-entropy rates of groups a and a', summed over all pairs of different groups:

$$\mathcal{L}_{EAR} = \frac{1}{2} \sum_{a,a'} \left| \ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a') \right| + C(\theta, a, a'), \tag{10}$$

where $C(\theta, a, a')$ is an offset term to be described shortly.

We do not have $p_{\hat{Y}|A}(Y|a)$, but we can estimate it based on the portion of the training data spoken by people from group a, thus

$$\ln p_{\hat{Y}|A}(Y|a) \approx \frac{1}{|S_a|} \sum_{x^{(i)},y^{(i)} \in S_a} \ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \tag{11}$$

where $|S_a|$ is the number of training utterances available from group a. The log-probability $p_{\hat{Y}|X}(y^{(i)}|x^{(i)})$ in Eq. (11) is no more than the negative cross-entropy loss of the training pair $(x^{(i)}, y^{(i)})$.

There are two possibilities to optimize $\mathcal{L}_{EAR}$, either increasing the performance of the group with lower accuracy or decreasing the performance of the one with higher accuracy. Apparently, the latter situation is not desirable. In order to avoid the latter, we can modify the equal accuracy ratio by adding an offset term, equal to the average of the two group-dependent cross entropies:

$$C(\theta, a, a') = -\frac{\ln p_{\hat{Y}|A}(Y|a) + \ln p_{\hat{Y}|A}(Y|a')}{2} \tag{12}$$

With the offset constant defined in Eq. (12), the equal accuracy ratio becomes a weighted average of the per-group cross-entropy losses:

$$\begin{aligned}\mathcal{L}_{EAR} &= \sum_{a,a'} \max\left\{-\ln p_{\hat{Y}|A}(Y|a), -\ln p_{\hat{Y}|A}(Y|a')\right\}, \\ &= -\sum_{a} N_{\leq a} \ln p_{\hat{Y}|A}(Y|a),\end{aligned} \tag{13}$$

where $N_{\leq a}$ is the number of other groups that have lower cross-entropy loss than group a. The resulting $\mathcal{L}_{EAR}$ as a measure of inclusiveness is intuitive. It is a weighted sum of cross-entropy loss over each dialect where the dialects with larger loss are given larger weights during training.

The loss function $\mathcal{L}_{CTC}$ penalizes high average error rates, but ignores high inter-group error rate disparities; $\mathcal{L}_{EAR}$ penalizes high inter-group disparities, but ignores the error rate of the best-performing system. Multi-task training seeks to balance these two objectives by minimizing

$$\mathcal{L}_{MT} = \mathcal{L}_{CTC} + \lambda \mathcal{L}_{EAR}, \tag{14}$$

where λ is a hyperparameter that can be tuned.

4. Datasets

In order to better study the inclusiveness of ASR, a new benchmark dataset was created, by combining existing datasets that cover seven racially, ethnically, and geographically diverse dialects of English. The new multi-dialect dataset is described in Section 4.1. The equal accuracy ratio is quite a general training criterion, however; its generality was also tested using two existing corpora, CORAAL and UASpeech, that have other types of protected attributes labeled as metadata. The additional protected attributes labeled in the CORAAL and UASpeech corpora are described in Section 4.2.

4.1. Multi-dialect dataset

We study the balance of error rates among speakers of different dialects by collecting data from seven published corpora, listed in Table 1. Most are retrieved from publicly available sources. Names for the American dialects (Standard American and African American) are based on the discussion in Wolfram and Schilling (2015). The Afrikaans English and Xhosa English corpora are named as in their source distribution (Roux et al., 2004); the Latin American, UK Broadcast News and Indian English corpora are each named for the country or countries in which they were recorded.

The African American corpus is part of the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2018), which provides recorded conversational speech data from people who self-identify as African American, including audio recordings, time-aligned orthographic transcription and speaker information. We use the DCA version[1] which focuses on African American Language in the Washington DC region. We omit speech by the interviewers and retain only speech segments by the self-identified speakers of African American Language.

The Standard American English corpus is collected from the LibriSpeech ASR[2] corpus (Panayotov et al., 2015) of audiobook readings. Data from this corpus are not dialect-homogeneous: all regional dialects of the United States are represented, as are samples of dialects from outside the United States. Nevertheless, empirical results reported later in this paper suggest that this corpus is more dialect-homogeneous than any of our other purportedly single-dialect corpora, possibly because speakers modulate their speech, somewhat, to match a standard audiobook reading style. The data contains audio waveforms and their associated transcriptions. Librispeech is a very large corpus; we use only the "train-clean-100" partition.

The Latin American English corpus is extracted from Hispanic-English Database (LDC2014S05)[3] (Byrne et al., 2014) that contains a mixture of read speech and conversational speech along with their transcriptions. Participants were adult native speakers of Spanish as spoken in Central America and South America who resided in the Palo Alto, California area, had lived in the United States for at least one year and demonstrated a basic ability to understand, read and speak English. We only include the read speech part of the corpus.

UK Broadcast News is extracted from WSJCAM0 Cambridge Read News (LDC95S24)[4] corpus (Robinson et al., 1995). The subjects in WSJCAM0 were native speakers of British English, reading in a standardized dialect. The corpus provides standard orthographic transcripts as well as time alignment between waveform and both word and phonetic transcriptions. The audio is originally in NIST SPHERE format and we convert it to wav format for fast data loading.

The AST Afrikaans English corpus and AST Xhosa English[5] (Roux et al., 2004) are collected and published by African Speech Technology.[6] AST Afrikaans English corpus contains a mixture of spontaneous and read speech by native speakers of Afrikaans, a language primarily spoken by white South Africans. AST Xhosa English is collected in a similar form but is spoken by native speakers of isiXhosa, a language primarily spoken by black South Africans. Both corpora are distributed in 8 kHz alaw format, and were converted to 16-bit 16kHz wav files using `FFmpeg`.

The Indian English corpus is composed of three small corpora posted to Voxforge[7] by Mahesh Chandra. Unlike other dialects, this corpus contains the speech of only one speaker, reading short sentences.

These corpora vary considerably in difficulty. The Indian English and Latin American corpora are difficult to recognize because they are small, and because the sampled dialects are quite different from the others in the list. The African American corpus is difficult because it is composed exclusively of spontaneous speech. The Afrikaans English and Xhosa English corpora each contain a mixture of spontaneous and read speech; the Standard American and UK Broadcast News corpora each contain exclusively read speech.

Transcriptions are cleaned by removing special characters and punctuation except apostrophe. We retain audio files with a duration longer than 1 second. Short-time Fourier transform (STFT) is computed using a 16 kHz sampling rate, and a Hamming window with a window size of 0.02 s and a window stride of 0.01 s. ASR features consist of the natural logarithm of one plus the magnitude of STFT, normalized by subtracting the mean and dividing by standard deviation.

[1] CORAAL: https://oraal.uoregon.edu/coraal.
[2] LibriSpeech: http://www.openslr.org/12.
[3] LDC2014S05: https://www.ldc.upenn.edu.
[4] LDC95S24: https://www.ldc.upenn.edu.
[5] AST Afrikaans English Corpus: https://vlo.clarin.eu/record/https_58__47__47_hdl.handle.net_47_20.500.12185_47_411_64_format_61_cmdi?2.
[6] AST Xhosa English is a subset of the AST Black English Corpus: https://repo.sadilar.org/handle/20.500.12185/433.
[7] Mahesh-Chandra Corpus: http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/8kHz_16bit.

Table 1
Sources of data used in our multi-dialect dataset "Abbr" column is the abbreviated dialect name used in performance tables. "#Utts" column shows the number of utterances in the training set. "Len" column shows the total duration of all utterances, in minutes.

Dialect	Abbr	Corpus	# Utts	Len
African American	AA	CORAAL	13908	491
Afrikaans Eng	AF	AST Afrikaans	3799	133
Standard American	AM	Librispeech	28533	6035
UK Broadcast News	BR	LDC95S24	10980	1221
Latin American	LA	LDC2014S05	281	28
Indian Eng	IN	MaheshChandra	358	16
Xhosa Eng	XH	AST Black	3323	116

Table 2
Partition of CORAAL dataset. "Abbr" column shows the abbreviated group name used in performance tables. "#Utts" column shows the number of utterance in the training set. "Len" column shows the total duration in minutes of the utterances.

Attr	Group	Abbr	#Utts	Len
Age	−19		7320	250
	20-29		2776	104
	30-50		2590	99
	51+		1122	37
Work	Lower Working Class	LW	3516	125
	Upper Working Class	UW	4359	146
	Lower Middle Class	LM	3647	131
	Upper Middle Class	UM	1159	46
	Upper Class	U	824	28
	Unknown.	Unk	403	13
Edu	Elementary School	ES	169	6
	Student in Middle School	StMS	3190	107
	Student in High School	StHS	3510	118
	Some High School.	SHS	1206	41
	High School	HS	3156	108
	Student in College	StCO	192	7
	Some College	SCO	1485	63
	College	CO	847	32
	Graduate School	GS	153	5
Gender	Male	M	9155	317
	Female	F	4753	174

4.2. CORAAL and UASpeech

We study the application of inclusiveness to attributes other than dialect using two corpora that contain other types of metadata labels: CORAAL, and UASpeech.

The CORAAL dataset is part of the multi-dialect dataset, and has been described in Section 4.1 (Kendall and Farrington, 2018). It contains complete and detailed information about both interviewers and interviewees. We extract interviewee information and identify 4 attributes, namely `age`, `work`, `education`, and `gender`, that are complete and meaningful to be used as sensitive attributes to partition the corpus. Details of the attributes are provided in Table 2. Additional experiments are performed on this dataset to verify the effectiveness of the equal accuracy ratio.

We study the application of inclusiveness to speakers with and without speech disabilities by using the UASpeech corpus[8] (Kim et al., 2008). The dataset contains speech from speakers with Cerebral Palsy, who self-report a diagnosis of dysarthria (reduced intelligibility caused by neuromotor disorder). Speech materials consist of 765 isolated words per speaker: 300 distinct uncommon words, and 3 repetitions each of digits, computer commands, radio alphabet and common words. The corpus is distributed with intelligibility ratings for each disabled speaker, calculated by asking human annotators to attempt to transcribe each person's speech: 0 – 25% intelligibility is called "Very Low", 25 – 50% is called "Low", 50 – 75% is called "Mid", 75 – 100% is called "High". From these data, we selected a control group (13 speakers without dysarthria) and an experimental group (8 speakers with dysarthria). UASpeech includes 15 speakers with dysarthria, but we choose to exclude speakers with Low or Very Low intelligibility (less than 50% of their isolated words are intelligible). These speakers exclude and repeat phones and entire syllables, therefore it is not always clear what a speech recognizer should transcribe.

Stationary noise in the recordings is first removed using `noisereduce` (Sainburg, 2019). Silence is trimmed from the beginning and end of each recording. Mel log spectrogram features are extracted using `librosa` (McFee et al., 2015), with a window stride of 0.01 s. ASR features consist of the natural logarithm of the magnitude of the mel filterbank features, normalized so that anything from $-\infty$ to -120 db is set to zero and 0 db is set to one.

5. Experimental methods

The new multi-dialect corpus is larger than CORAAL, which is, in turn, larger than UASpeech, therefore it is inappropriate to use equal hyperparameters (equal numbers of layers, equal number of nodes, equal selection of layer types) for all three datasets. Architectures were therefore tuned to each dataset, in order to optimize the accuracy of the baseline system. Our code is available online.[9]

5.1. Multi-dialect dataset: Methods

The multi-dialect dataset was split into train, dev, and test subcorpora with a ratio of 8: 1: 1. The training subcorpus was used to train a `deepspeech` network[10] (Hannun et al., 2014). The deepspeech model has 2 convolutional layers, each with batch normalization and tanh activation. The convolution kernel sizes are 41×11 and 21×11 respectively. The convolution output is passed to 5 batch-normalized bidirectional LSTM layers, whose output is fed into a fully connected layer. The output is softmaxed to predict a label distribution at each frame, which is then used to the multi-task training loss $\mathcal{L}_{MT}$ in Eq. (14).

The equal accuracy ratio requires weighting the dialects in ascending order of their average CTC loss terms, as shown in Eq. (13). Ideally, the average CTC loss for each dialect should be calculated in a large batch containing all utterances. Due to GPU memory limitations, our model can only be trained with a batch size of 16. In order to determine dialect weightings, therefore, the average CTC loss of each dialect is calculated cumulatively with new incoming batches within an epoch, so as the training proceeds, the estimated average CTC loss and equal accuracy ratio become increasingly accurate. The model is optimized using Adam optimizer with a learning rate of 0.001. Each model is trained for 30 epochs and early-stopped based on validation loss. Models are evaluated on the test dataset with character error rate (CER) as the evaluation metric.

The equal accuracy ratio (as in Eq. (13)) has no built-in dependence on the historical or socioeconomic importance of the group definitions; it simply penalizes the largest group loss. For example, if it is desirable that every single utterance should have exactly the same error rate, EAR can be re-designed in pursuit of that goal: the maximization in Eq. (13) is simply computed for every pair of utterances, regardless of their group affiliations. We tested this per-utterance version of EAR on the dialect dataset. This experiment uses the same loss functions as shown in Eqs. (13) and (14), but instead of assigning greater weight to the dialects with higher loss, we assign greater weight to the utterances with higher loss.

[8] UASpeech: http://isle.illinois.edu/sst/data/UASpeech/index.html.

[9] Code: https://github.com/Hertin/Equal-Accuracy-Ratio.

[10] Deepspeech: https://github.com/SeanNaren/deepspeech.pytorch.

Table 3
Character error rate (CER: percent) and word error rate (WER: percent), measured as a function of dialect and of the regularization weight λ, in experiments using the multi-dialect corpus. WERs are in the parentheses. Refer to Table 1 for the meanings of the dialect abbreviations. Mean and Std are the mean and standard deviation across dialects; Mean-IU and Std-IU are the mean and standard deviation across utterances..

CER (WER)	Multitask Regularization Weight λ, Inter-Group (IG) Regularization					
Dialect	0	0.001	0.01	0.1	1	10
AA	43.38 (82.37)	**39.30** (**76.47**)	41.88 (79.20)	43.52 (82.62)	42.28 (78.35)	46.16 (85.11)
AF	16.20 (35.47)	**14.01** (**30.60**)	15.99 (35.88)	16.50 (35.32)	15.94 (33.16)	18.16 (40.54)
AM	13.61 (39.51)	**10.24** (**31.02**)	12.27 (35.99)	13.62 (40.07)	12.55 (37.68)	16.01 (45.65)
BR	13.11 (41.00)	**10.05** (**33.09**)	11.93 (37.60)	13.09 (41.93)	12.72 (42.19)	16.52 (50.47)
IN	51.65 (89.38)	**50.40** (**89.01**)	51.54 (90.64)	52.16 (89.18)	50.91 (91.52)	53.64 (95.91)
LA	41.04 (86.09)	**32.52** (**75.56**)	34.08 (79.83)	35.97 (83.83)	32.93 (75.94)	38.81 (87.22)
XH	25.64 (52.46)	**21.04** (**45.52**)	23.40 (48.25)	23.75 (50.21)	22.97 (47.24)	25.91 (55.07)
Mean	29.23 (60.90)	**25.37** (**54.47**)	27.30 (58.20)	28.37 (60.45)	27.19 (58.01)	30.74 (65.71)
Std	14.78 (22.29)	14.56 (23.23)	14.40 (22.28)	14.47 (21.90)	**14.10** (21.56)	14.28 (**21.15**)
Mean-IU	22.58 (52.92)	**19.92** (**47.58**)	21.29 (49.95)	22.49 (53.53)	21.68 (50.76)	25.24 (58.75)
Std-IU	17.92 (31.09)	18.33 (32.43)	**17.87** (31.73)	17.92 (31.35)	17.94 (**30.03**)	18.16 (30.91)
CER (WER)	Multitask Regularization Weight λ, Inter-Utterance (IU) Regularization					
Dialect	0	0.001	0.01	0.1	1	10
AA	43.38 (82.37)	42.48 (81.55)	**41.31** (**76.43**)	43.40 (82.16)	41.50 (79.61)	44.07 (83.34)
AF	16.20 (35.47)	16.88 (36.44)	**14.95** (**29.73**)	16.58 (36.34)	15.21 (33.93)	18.32 (42.08)
AM	13.61 (39.51)	13.94 (39.55)	**10.62** (**32.14**)	13.99 (39.80)	12.86 (35.88)	16.67 (47.33)
BR	13.11 (41.00)	11.06 (40.37)	**10.40** (**34.20**)	13.29 (40.93)	11.68 (38.28)	16.71 (51.16)
IN	51.65 (**89.38**)	52.91 (92.40)	52.08 (93.57)	51.94 (91.81)	52.16 (92.40)	**51.25** (93.86)
LA	41.04 (86.09)	38.00 (85.71)	31.98 (**73.68**)	37.12 (86.09)	**29.55** (75.56)	38.54 (86.84)
XH	25.64 (52.46)	24.88 (50.33)	**21.29** (**45.05**)	24.89 (51.81)	23.09 (49.91)	25.93 (57.02)
Mean	29.23 (60.90)	28.59 (60.91)	**26.09** (**54.97**)	28.74 (61.28)	26.58 (57.94)	30.21 (65.95)
Std	14.78 (22.29)	14.86 (22.75)	15.00 (23.87)	14.35 (22.59)	14.33 (22.30)	**13.25** (**19.76**)
Mean-IU	22.58 (52.92)	21.98 (52.59)	**19.27** (**46.08**)	22.82 (52.15)	20.86 (49.90)	24.86 (57.22)
Std-IU	17.92 (31.09)	17.90 (31.22)	17.12 (30.93)	17.87 (31.39)	17.64 (30.54)	**16.77** (**29.29**)

5.2. CORAAL dataset: Methods

The CORAAL corpus is too small to train a complete deepspeech network; initial experiments overfit the training dataset, producing unstable performance on the development dataset. Since our equal accuracy ratio regularization does not depend on any specific neural architecture, we turn to a much simpler RNN-CTC architecture for a more stable performance measure. The model is composed of 4 layers of bidirectional LSTM with 128-dimensional hidden states, a batch normalization layer, and 4 fully connected layers, each with `tanh` activation. The output of the last layer is softmaxed and is used to compute CTC loss. Due to the simple RNN-CTC architecture, we are able to increase the batch size to 32 in training. Other settings are the same as those of dialect experiments.

The CORAAL corpus contains metadata about each interviewee, including `age`, `work`, `education`, and `gender` (Table 2). We ran four sets of experiments using this corpus, each of which used one of the metadata variables as an attribute protected by the equal accuracy ratio training criterion.

5.3. UASpeech dataset: Methods

UASpeech utterances are isolated words. Instead of character error rate, therefore, experiments with UASpeech report word error rate. Empirically, it was found that small word error rate reductions were achieved by converting each word into a phone sequence using LanguageNet grapheme-to-phoneme transducers (Hasegawa-Johnson et al., 2020), training an ASR to produce phone sequences as output, and then counting the word to be correct only if all of its phones were correctly recognized.

Experiments on the UASpeech corpus use a 4-layer bi-directional LSTM with a hidden size of 200 and a dropout of 0.1. The output is then passed through two fully-connected layers with a hidden layer size of 500.

We apply the `tanh` nonlinearity to the output of the first fully-connected layer and softmax to the output of the second fully-connected layer. The resulting logits are used to compute CTC Loss. The model is trained with a batch size of 128.

6. Experimental results

Multi-dialect recognition experiments are tested with different λ values in the range $0 \leq \lambda \leq 10$, where $\lambda = 0$ is the baseline, and $\lambda = 1$ gives equal weight to the cross-entropy and equal accuracy ratio loss terms, as shown in Eq. (14). Resulting character error rates (CER: percent) and word error rates (WER:percent) are listed in Table 3. `Mean` CER (WER) is the average per-dialect character error rate, averaged uniformly over the seven dialect groups. The standard deviation (`Std`) of CER (WER) over all dialect groups is a measure of inclusiveness. In addition, we compute the inter-utterance mean (`Mean-IU`) of CER (WER) by computing the average of per-utterance CER (WER) and inter-utterance standard deviation (`Std-IU`) of CER (WER) by computing the standard deviation of CER over all utterances.

In general, we find that increased emphasis on the equal accuracy ratio (higher λ) results in a lower standard deviation of the CER (increased inclusiveness). Optimal inclusiveness (minimum inter-group standard deviation) is achieved with a value of $\lambda = 1$. Optimal accuracy (minimum inter-group average error rate) is achieved with a value of $\lambda = 0.001$.

Table 4 provides experimental results from the CORAAL corpus. Four sets of experiments were conducted using this dataset, each treating a different metavariable as the protected attribute: `age`, `work`, `education`, and `gender`. The figures in the table are character error rates (CER: percent). Optimal inclusiveness (minimum inter-group standard deviation) is achieved for different values of λ, depending on the metavariable being protected. According to the empirical results, variation across different educational groups is minimized with $\lambda = 1$, variation across age groups is minimized with $\lambda = 0.1$, and variation across work or gender is minimized with $\lambda = 0.01$. Similarly, optimal average accuracy (minimal inter-group average error rate) is achieved for different values of λ, ranging from $\lambda = 0.001$ (`age` and `education`) to $\lambda = 10$ (`gender`).

Experiments on the UASpeech corpus are performed using three values of λ (0, 0.33 and 1) where $\lambda = 0$ is the base case without regularization. Results are reported in Table 5. The UASpeech corpus consists of isolated words, therefore results are reported in terms

Table 4
Character error rate (CER: percent), measured as a function of dialect and of the regularization weight λ, in experiments using CORAAL. Refer to Table 2 for the meanings of the group abbreviations.

	Multitask Regularization Weight λ					
Age	0	0.001	0.01	0.1	1	10
–19	55.59	56.60	**53.96**	56.23	55.94	56.72
20–30	55.56	55.99	**53.73**	55.82	56.60	57.13
30–50	56.31	56.99	**54.94**	56.24	56.61	57.04
50+	59.31	59.97	58.59	**58.53**	59.33	59.79
Mean	56.69	57.39	**55.30**	56.70	57.12	57.67
Std	1.78	1.77	2.25	**1.23**	1.50	1.42
Work						
LM	56.16	**54.97**	58.03	55.64	57.05	56.90
LW	55.30	**54.30**	57.44	55.06	56.76	55.60
UW	56.03	**54.68**	58.32	55.55	56.96	56.81
UM	58.01	**55.62**	58.27	55.69	58.15	57.78
U	58.76	**57.25**	59.06	57.33	59.31	57.99
Unk	56.86	**54.71**	57.41	57.46	56.71	56.36
Mean	56.85	**55.26**	58.09	56.12	57.49	56.91
Std	1.31	1.07	**0.62**	1.01	1.04	0.89
Edu						
ES	61.94	61.00	61.54	62.35	**59.24**	60.19
StMS	55.54	**54.86**	55.95	57.03	57.28	56.93
StHS	55.40	**54.55**	56.48	57.31	56.71	55.83
SHS	**55.20**	55.25	56.70	57.73	56.87	55.57
HS	57.27	**56.04**	58.63	59.13	58.06	56.69
StCO	**51.95**	53.25	55.03	59.17	54.79	57.28
SCO	56.12	**55.54**	57.27	57.99	57.48	56.65
CO	54.18	**53.79**	55.70	55.62	55.28	55.04
GS	**54.42**	54.97	54.83	57.04	56.22	55.39
Mean	55.78	**55.47**	56.90	58.15	56.88	56.62
Std	2.74	2.24	2.09	1.92	**1.36**	1.54
Gender						
M	55.74	55.28	55.55	57.32	58.07	**55.21**
F	55.93	56.41	55.56	57.57	57.46	**55.44**
Mean	55.84	55.85	55.55	57.45	57.76	**55.32**
Std	0.13	0.80	**0.01**	0.17	0.43	0.16

Table 5
Word error rate (WER: percent), measured as a function of group identity and of the regularization weight λ, in experiments using UASpeech. Control group: speakers without dysarthria, Exp Group: speakers with Dysarthria.

UASpeech	$\lambda = 0$	$\lambda = 0.33$	$\lambda = 1$
Control	2.84	2.54	**2.45**
Exp	6.28	5.08	**4.52**
Mean	4.56	3.81	**3.49**
Std	2.43	1.79	**1.46**

of word error rate (WER: percent) rather than character error rate. Optimal inclusiveness (minimum inter-group standard deviation) and optimal average accuracy (minimum inter-group average error rate) are both achieved with a value of $\lambda = 1$.

7. Discussion

By comparing the CER and WER in multi-dialect experiments, we observe that these two metrics show similar dependence on λ: optimum values of `Mean` and `Mean-IU` are always achieved for the same values of λ, while optimum values of `Std` and `Std-IU` are achieved for similar values of λ. None of these measures are convex functions of λ, but measures based on CER tend to be closer to convex than measures based on WER. The extreme non-convexity of WER measures may be caused by the unpredictable way in which the deepspeech model represents language model information. Our deepspeech implementation has no external language model; its only language model is a character-based language model, learned implicitly by the LSTM layers. Differences in word choice and word use among these seven dialects may not be well-represented by a character-level language model, resulting in erratic behavior of the model as a function of λ.

Experiments verify that with the regularization, we are able to achieve a more inclusive model. Dialect experiments in Table 3, CORAAL experiments in Table 4 and UASpeech experiments in Table 5 all show a consistent reduction in standard deviation when using the equal accuracy ratio term as a regularizer. Because of the regularization, the groups that have the worst performance, during each epoch of training, are assigned more weight in the training. In experiments on the multi-dialect corpus, the most difficult dialect (Indian English) is improved more than other dialects. Similar findings are observed for the experimental group in the UASpeech experiments.

By comparing the inter-dialect (Dialect) experiment results with the inter-utterance (Dialect-IU) results, we observe that, giving greater weight to the most poorly served dialect uniformly reduces the inter-dialect standard deviation (`Std`); inter-utterance standard deviation (`Std-IU`) sometimes drops, but not always. On the other hand, assigning greater weight to the most poorly served utterance uniformly reduces `Std-IU`; `Std` sometimes drops, but not always.

There is usually a trade-off between accuracy and inclusivity: average CER (accuracy) and standard deviation of CER (inclusivity) are optimized at different values of λ. The trade-off can be observed in the dialect experiments in Table 3, and in the CORAAL experiments in Table 4. We observe that assigning more weight (λ) to the equal accuracy ratio results in smaller standard deviation figures and thus more inclusiveness of the system. However, at the same time, larger λ also tends to increase the mean error rate and reduce the accuracy of the system.

Training using the equal accuracy ratio does not always harm average accuracy, however; for very small values of λ, multi-task training, using the equal accuracy ratio as a regularizer, may actually improve the average accuracy of the system, averaged across both advantaged and disadvantaged groups. Non-zero values of λ yield the lowest average CER for all of the experimentally tested group attributes, including dialect (Table 3), `age`, `work`, `education` and `gender` (Table 4), and disability (Table 5). Possibly the regularization forces the model to pay more attention to boundary cases (high loss cases) and thus the resulting model can have better generalization during testing.

8. Conclusions

This paper defines a new multi-dialect corpus that can be used to study the inclusiveness of automatic speech recognizers. A new training criterion designed to enhance inclusiveness by minimizing the maximum error rate is proposed for variable-length sequence generation problems. Since sequence scoring is usually done in the log probability domain, the proposed inclusiveness criterion is based on an accuracy ratio, rather than an accuracy difference; since sequence scoring problems are trained without known time alignment, the proposed criterion is marginalized over all possible time alignments. Experiments are performed to demonstrate that the proposed equal accuracy ratio can be seamlessly integrated into CTC-based ASR neural architecture as a regularizer to improve the inclusiveness of the ASR model. Several sets of experiments are performed to verify the effectiveness of the proposed method. Results demonstrate that the models with inclusiveness regularization generally have a smaller performance gap among user groups, without increasing the overall average error rate, as compared to baseline.

CRediT authorship contribution statement

Heting Gao: Methodology, Writing – original draft, Software. **Xiaoxuan Wang:** Writing – original draft, Software, Data curation. **Sunghun Kang:** Conceptualization, Writing – review & editing. **Rusty Mina:** Conceptualization, Writing – review & editing. **Dias Issa:** Conceptualization, Writing – review & editing. **John Harvill:** Conceptualization, Data curation, Writing – review & editing. **Leda Sari:** Conceptualization, Writing – review & editing. **Mark Hasegawa-Johnson:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Chang D. Yoo:** Conceptualization, Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.specom.2021.11.004. This work was supported by ICT.

References

Anahideh, H., Asudeh, A., 2020. Fair active learning. arXiv preprint arXiv:2001.01796.

Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N., 2016. Lipnet: Sentence-level lipreading. 2, (8), arXiv preprint arXiv:1611.01599.

Byrne, W., et al., 2014. Hispanic–English database LDC2014S05. DVD.

Calders, T., Kamiran, F., Pechenizkiy, M., 2009. Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops. IEEE, pp. 13–18.

Carter, P.M., Callesano, S., 2018. The social meaning of Spanish in Miami: Dialect perceptions and implications for socioeconomic class, income, and employment. Latino Stud. 16, 65–90.

Désilets, A., Stojanovic, M., Lapointe, J.-F., Rose, R., Reddy, A., 2008. Evaluating productivity gains of hybrid ASR-MT systems for translation dictation. In: IWSLT. pp. 158–165.

Elfeky, M.G., Moreno, P., Soto, V., 2018. Multi-dialectical languages effect on speech recognition: Too much choice can hurt. Procedia Comput. Sci. 128, 1–8.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Internat. Conf. Machine Learning (ICML). pp. 369–376. http://dx.doi.org/10.1145/1143844.1143891.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al., 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. pp. 3315–3323.

Hasegawa-Johnson, M., Rolston, L., Goudeseune, C., Levow, G.-A., Kirchhoff, K., 2020. Grapheme-to-phoneme transduction for cross-language ASR. Lecture Notes in Comput. Sci. 12379, 3–19.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., Itahashi, S., 1999. Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Japan (E) 20 (3), 199–206.

Kamiran, F., Calders, T., 2009. Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication, pp. 1–6.

Kendall, T., Farrington, C., 2018. The corpus of regional african american language. Version 6, 1.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1741–1744.

Kleinberg, J., Mullainathan, S., Raghavan, M., 2017. Inherent trade-offs in the fair determination of risk scores. In: Papadimitriou, C.H. (Ed.), 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). In: Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, (ISSN: 1868-8969) ISBN: 978-3-95977-029-3, pp. 43:1–43:23. http://dx.doi.org/10.4230/LIPIcs.ITCS.2017.43, URL http://drops.dagstuhl.de/opus/volltexte/2017/8156.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S., 2020. Racial disparities in automated speech recognition. Proc. Natl. Acad. Sci. 117 (14), 7684–7689.

Lamel, L.F., Kassel, R.H., Seneff, S., 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Proc. of the DARPA Speech Recognition Workshop, pp. 100–109.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J.M., Nguyen, H., Gadde, R.T., 2019. Jasper: an end-to-end convolutional neural acoustic model. In: Proc. Interspeech, pp. 71–75.

Li, B., Sainath, T.N., Sim, K.C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., Rao, K., 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4749–4753.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O., 2015. librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp. 18–25.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5206–5210.

Picone, J., 1990. The demographics of speaker independent digit recognition. In: Proc. ICASSP, pp. 105–108.

Robinson, T., et al., 1995. WSJCAM0 Cambridge Read News (LDC95S24). Linguistic Data Consortium, Philadelphia, PA.

Roux, J.C., Louw, P.H., Niesler, T., 2004. The african speech technology project: An assessment. In: Language Resources and Evaluation Conference LREC. European Language Resources Association ELRA, pp. 93–96.

Sainburg, T., 2019. Timsainb/noisereduce: Initial release. Zenodo URL https://zenodo.org/record/2596682#.X1ekq8gzZPY.

Scharenborg, O., 2021. Inclusive speech technology: Developing automatic speech recognition for everyone. Webinar delivered to the TU Delft Safety and Security Institute and Campus, The Hague, The Netherlands.

Schlüter, R., 2019. Survey talk: Modeling in automatic speech recognition: Beyond hidden Markov models. Survey talk presented at Interspeech 2019.

Shillingford, B., Assael, Y.M., Hoffman, M.W., Paine, T., Hughes, C., Prabhu, U., Liao, H., Sak, H., Rao, K., Bennett, L., Mulville, M., Coppin, B., Laurie, B., Senior, A.W., de Freitas, N., 2018. Large-scale visual speech recognition. CoRR http://arxiv.org/abs/1807.05162.

Tatman, R., 2017. Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 53–59.

Vipperla, R., 2011. Automatic speech recognition for ageing voices. (Ph.D. thesis). The University of Edinburgh.

Winata, G.I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., Fung, P., 2020. Learning fast adaptation on cross-accented speech recognition. arXiv preprint arXiv:2003.01901.

Wolfram, W., Schilling, N., 2015. American English: Dialects and Variation. Jon Wiley & Sons, New York.

Yang, X., Audhkhasi, K., Rosenberg, A., Thomas, S., Ramabhadran, B., Hasegawa-Johnson, M., 2018. Joint modeling of accents and acoustics for multi-accent speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.

Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P., 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1171–1180.

Zeyer, A., Bahar, P., Irie, K., Schlüter, R., Ney, H., 2019. A comparison of transformer and LSTM encoder decoder models for ASR. In: Proc. ICASSP, pp. 8–15.