Analyzing the Quality and Stability of a Streaming End-to-End On-Device Speech Recognizer

Yuan Shangguan†, Kate Knister*, Yanzhang He, Ian McGraw, Françoise Beaufays*

Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA

yuansg@fb.com,kateknister@google.com

Abstract

The demand for fast and accurate incremental speech recognition increases as the applications of automatic speech recognition (ASR) proliferate. Incremental speech recognizers output chunks of partially recognized words while the user is still talking. Partial results can be revised before the ASR finalizes its hypothesis, causing instability issues. We analyze the quality and stability of on-device streaming end-to-end (E2E) ASR models. We first introduce a novel set of metrics that quantify the instability at word and segment levels. We study the impact of several model training techniques that improve E2E model qualities but degrade model stability. We categorize the causes of instability and explore various solutions to mitigate them in a streaming E2E ASR system.

Index Terms: ASR, stability, end-to-end, text normalization, on-device, RNN-T

1. Introduction

Modern applications of automatic speech recognition have brought to users fast and accurate incremental speech recognition experiences. These speech recognizers stream chunks of partially recognized words to the user interface while the user is still talking. We hereby refer to these chunks as *partials*. Not only do users see text appear in real-time before the ASR recognizer finalizes the transcription, downstream models such as spoken dialog systems [1], real-time translation system [2] and multimodal user interfaces [3] also rely on partials to reduce overall application latencies.

Prior works in streaming ASR highlight a significant obstacle: model stability [1, 4, 5]. Recognizers often change emitted words before finalizing the hypothesis. These revisions of partials cause flicker on the speech user interface. They also increase the cognitive load of the users, distracting them from speaking and resulting in frustration. Revisions also cause the downstream models to repeat operations, increasing overall application latency.

In this paper, we analyze the stability issues of an on-device, streaming, Recurrent Neural Network Transducer (RNN-T) ASR recognizer. It is a state-of-the-art E2E ASR recognizer that produces transcripts with low word error rates (WER) on devices while staying within the constraints of latency, memory storage and computational resources [6]. We improve the training data diversity by mixing data from multiple speech domains [7, 8]. Multi-domain training data improves robustness of the model for conditions not seen during training, such as long-form audio recognition [9]. In this paper, we analyze the impact of the diverse training data on the quality and stability of RNN-T recognizers. Moreover, to avoid the additional latency from text normalizers, we include capitalizations, spoken punctuations, and Arabic numerals in the training data, so that the trained models output these written-text formats [10]. Due to the diversity of transcript text formats present in the multi-domain training data, the on-device RNN-T is especially prone to instability, revising its transcripts frequently during the generation of partials.

Prior work proposed different methods for improving and evaluating stability of traditional streaming ASR recognizers. Both [5] and [11] defined a stable partial as an unchanging prefix that prepends the growing hypothesis. They modeled the stability statistics of partial hypotheses using logistic regression or a single-hidden-layer feedforward network and suggested emitting only partials that are classified as stable. In constrast, Selfridge et al. [1] defined partial prefixes as stable only if they prefix the final transcription hypothesis. They reported the percentage of stable partials by generating partials only at certain nodes in the lattice. Baumann [4] measured the cumulative differences between subsequent partials up to the hypothesis. When users see incremental changes in transcription, the cumulative measurement provides better insight into user experience. In a similar vein, a recent paper in speech-to-translation [2] used erasure. Erasure measures the number of tokens to be removed from the suffix of the previous translation partials to produce the next sequence of partials. The ratio between aggregated erasure and the final translation length measures instability of the translation system.

To examine the stability of on-device streaming E2E ASR models, we propose a novel **set** of metrics in Section 2 that explicitly measure the speech recognizer stability users perceive. These metrics are simple and intuitive. They can be captured live without incurring extra latency on user devices. More importantly, this set reflects both the frequency and span of revisions in the partials of a speech recognizer. We then analzye the quality, latency and stability of a RNN-T based E2E streaming speech recognizer. We look at the impact of training techniques we used for E2E models on the stability of RNN-T recognizers: mixed-case data, multi-domain training data, and text normalization training. We categorize the causes of instability in Section 3. Most existing methods to improve ASR stability focus on delaying partials from the speech recognizer. We follow this trend to analyze the trade-offs between end-to-end speech recognizer's latencies and stabilty in Section 5. In addition, we evaluate our strategies in improving speech recognizer stability without delaying partial emissions in Section 4.

2. Instability Metrics

We illustrate the way we measure stability with the example in Table 1. In this contrived example, the user said "here comma lived a man who sailed to sea". The final recognizer output is "Here, lived a man who sailed to sea". A segment, or partial result, can contain multiple words. The emitted segments are

*Equal Contribution

†This work was done while author was working at Google

chronologically indexed. There are 9 segments in total, although the number could change if the frequency of partial emissions changes. A segment differs from the previous segment either by growth (addition of words) or by revision (changes in the previous words). Revisions indicate unstable segments. For example, segment 2 is a stable segment because it simply grows from segment 1. Segment 3, however, is an unstable segment because it contains a revision of the word "come" into "comma". Similar to [2, 5], we consider a partial prefix stable if and only if all future hypotheses contain the same prefix. In segment 5, for instance, "a man who" are unstable words because they follow an unstable word "Lived". As a result, we count 4 unstable words in segment 6.

Seg #	Streamed Segment Text	# Unstable	
		Word	Seg
1	Here		
2	Here come		
3	Here comma	+1	+1
4	Here,	+1	+1
5	Here, Lived a man who		
6	Here, lived a man who sell	+4	+1
7	Here, lived a man who sell two seeds		
8	Here, lived a man who sell 2 seeds	+2	+1
9	Here, lived a man who sailed to sea	+3	+1

Table 1: *Incremental speech recognition segments ordered by emission sequences, with unstable words and segment counts to calculate UPWR and UPSR.*

We measure the instability of partials with unstable partial word ratio (UPWR) and unstable partial segment ratio (UPSR). UPWR is the ratio of total number of unstable words in a test corpus to the total number of words in the final hypotheses. UPSR captures the ratio between the aggregated number of revised segments and the total number of utterances in a dataset. The ranges of UPWR and UPSR are $[0, \infty)$. The closer they are to 0, the more stable the system. UPWR captures the magnitude of the model's instability in terms of number of words, while the UPSR measures the frequency of occurences of revisions.

In Table 1, we have a total of 11 unstable words, and 9 words in the final hypothesis. Therefore UPWR=11/9=1.22. There are 5 unstable segments over one utterance, so UPSR=5/1=5.0.

3. Stability of the Streaming RNN-T

3.1. On-Device Streaming RNN-T

We train a streaming RNN-T with either 128 grapheme targets or 4096 word-pieces targets [10]. The RNN-T model is trained to handle multiple speech recognition domains [8]. In addition, this RNN-T is trained to run on modern edge devices [6]. Since on-device speech recognizers face serious constraints in latency, memory, and CPU resources, we avoid adding text normalizers to the speech pipeline. Instead, we include capitalizations, spoken punctuations and Arabic numerals in the training data, so that the trained RNN-T model learns to output transcripts with correct text normalization formats.

3.2. Types of On-Device RNN-T Instabilities

We analyze the streaming RNN-T output and divide the occurrences of instability into two categories. The first is text-normalization instability. The second is streaming instability. We further identify 4 subtypes of text-normalization instability. Table 2 shows the percentage breakdown of each instability subtype in terms of the frequency of occurrences in the output of the multi-domain RNN-T model, introduced in [9], which corresponds to Model F in Table 4.

Type of Stability	Percentage
A. Text Normalization Instabilities	**47.6%**
1,2. Punctuation & spacing related	21.2%
3. Capitalization	24.7%
4. Numeral	1.7%
B. Streaming instability	**52.4%**

Table 2: *Types of instability and their percentages of occurrences in an offline dataset.*

Streaming instability When the model is forced to output partial words faster, more premature partials occur when the user is still in the process of uttering a word. For example, one might see "my open" before "my opinion". We analyze the relationship between streaming instability and partial emissions rate in section 5. The streaming instability problem is more pronounced when the E2E models output subword targets instead of word targets as in the traditional models.

Text normalization instabilities arise when the RNN-T model revises its transcripts in terms of the formats of the output.

1. **Punctuation instability** refers to the change of partials related to spoken punctuation phrases. Examples include segment 2 to 4 in Table 1, where "come"→"comma" →"," are partial result revisions due to the comma symbol. Punctuation instability gets worse as the punctuation phrases get longer. "left quotation mark", for example, causes more instability than "period".

2. **Spacing instability** refers to the changes in spaces delimiting the partial words during the recognition process. It often happens hand-in-hand with punctuation instability. A commonly occurring observation is that the space between a word and the subsequent punctuation is being removed and re-inserted multiple times: "Hi," → "Hi ," → "Hi, ". Languages that do not require delimitation, such as Chinese and Japanese, have no spacing instabilities.

3. **Capitalization instability** is caused by the model revising uppercase outputs into lowercase outputs or vise-versa. Segments 5 to 6 in Table 1 are examples of capitalization instability (i.e. "Lived" → "lived"). Capitalization instability is more pronounced in languages like German, where nouns are capitalized.

4. **Numeral instability** occurs when users dictate phone numbers, street addresses, dates, time or other numeric entities. When the user is speaking, the ASR models output these numbers in spoken format first - "Call eight" - before outputting numbers - "Call 800-123-1234".

3.3. Factors Impacting Stability in RNN-T

We identify factors that adversely impact the stability of a mixed-case streaming RNN-T model in the following subsections.

3.3.1. Wordpieces and Spoken Punctuation Tokens

Word-piece targets in E2E speech models allow models to capture longer linguistic and phoneme contexts, resulting in better model performance [12]. We train a word-piece model (WPM) with n-gram counts obtained from text data to segment words into sub-word units. WPMs are depicted by Schuster in [13]. These WPM targets do not contain spoken punctuation tokens. For example, "exclamation" is represented by sub-word units as {"exclam", "a", "tion"}. As a result, the RNN-T models emit partial punctuation words like "exclam" before completely converting the spoken punctuations "exclamation mark" into "!".

3.3.2. Mixed-case Training Data and WPM Targets

In Section 3.1, we explain why we introduce mixed-case training data to RNN-T models. In terms of WER, RNN-T with WPM targets have lower WER than RNN-T with grapheme targets. In terms of stability, mixed-case training inevitably increases the perplexity of the RNN-T decoder. RNN-T models with WPM targets suffer greater stability degradations than RNN-T models with grapheme targets when the training data changes from lowercase to mixed-case. Table 3 shows stabilities of Model A (grapheme) and B (WPM) with lowercase traing data. They have similar stability. However, RNN-T Model D, trained with a mixed-case WPM, is 31.3% worse in stability than its grapheme-trained counterpart C. Comparing grapheme model pairs C and A, the grapheme-trained RNN-T has $> 40\%$ increase in UPWR and UPSR. Comparing the WPM models B and D, the RNN-T experiences a 61.5% increase in UPWR and 90.7% increase in UPSR.

One might surmise that mixed-case data mainly contributes to an increase in capitalization instability. Evidence suggests otherwise: mixed-case training causes other types of instability in WPM-target RNN-T. To show this, we post-process the outputs of models C and D with a lowercasing Finite State Transducer (FST). This FST converts all partial outputs into lowercase text, eliminating capitalization instability completely. The newly augmented models are *Cnorm* and *Dnorm* in Table 3. *Cnorm* has the same level of instability as Model A, suggesting that for grapheme RNN-T models, the $> 40\%$ increase in mixed-case instability is due to capitalization instabilities alone. *Dnorm*, however, has 30.2% higher UPSR than Model B. We find that the mixed-case decoding beams of the WPM-targeted RNN-T often contain more words with both upper and lower cases than the grapheme-targeted RNN-T. For example, "used" and "Used" can both exist in two otherwise identical top-N hypotheses because both are valid words. As a result, the diversity of the top N hypotheses decreases, and the top-1 hypothesis is more likely to change during the decoding process when the top-n hypotheses swap their rankings.

3.3.3. Multi-domain Training Data

Multi-domain training data improves the quality of the RNN-T model. It supplies a variety of acoustics and linguistic content to the speech model, improving model performance [9]. The cost of multi-domain training data is that the transcription standard may not be the same across all domains. Capitalizations, for instance, might not be enforced in the YouTube transcriptions. As the diversity of transcription format increases, multi-domain training leads to more model instability. In Table 4, we compare two models: D, trained with data from a single domain, and E, trained with data from multiple domains. Although multi-domain RNN-T model E has a better word error rate (WER) than model D, its punctuation error rate (PER) increases by 10.7%, case-insensitive WER increases by 1.5%, and its instability metrics increase even more drastically: UPWR $+76.2\%$ and UPSR $+45\%$.

4. Improving Text Normalization Stability

In this section, we discuss methods that improve E2E RNN-T recongizer stability without introducing a delay in partial emissions. We develop strategies to reap the benefits of improved WER while alleviating the degradations of model stability explained in section 3.3.

4.1. Numeral Instability

To reduce numeral instability, we included Text-to-speech (TTS) synthesized number data as introduced in [10] to fine-tune the models so that the RNN-T models automatically output the correct numeral formats.

4.2. Punctuation Words as WPM Tokens

To eliminate punctuation instabilities, we force the RNN-T models to predict punctuation phrases as single tokens. We add a list of possible punctuation phrases in the single-token format, such as "{exclamation-mark}" or "{left-curly-bracket}". We then add these tokens in the WPM vocabulary. At training time, we pre-process the audio transcriptions to ensure that dictated punctuations occur in the same token format. The trained ASR model learns to distinguish common words from punctuation phrases in their linguistic context; it predicts "period" when it means a stretch of time and "{period}" when it means end-of-sentence. At inference time, punctuation words with curly brackets are instantly converted into symbols by a simple regular expression logic.

4.3. Text Normalization and Domain-id

To alleviate the capitalization instability in mixed-cased trained RNN-T models, we implement two solutions: text normalization on the multi-domain dataset, and separation of domains using domain-id as a model input feature.

We apply text normalization models to the multi-domain training data [14] to unify the format of capitalization, spoken punctuations, numerics, and spacing clean-ups. Comparing Model E to F in Table 4, text normalization improves the stability of multi-domain model by 29.7% in UPWR and 19.3% in UPSR.

We use domain-id as input to the RNN-T model as explained in [8]. Domain-id allows the model to distinguish different text normalization standards and learn a more consistent text format from each domain. Model G with domain-id shows about 50% stability improvement over Model F.

With text normalization and domain-id, we develop Model G, a mixed-case streaming RNN-T with WPM targets. It has 23.4% better WER, 15.9% better case-insensitive WER, 8% better punctuation error rate and 12.5% improved UPWR than our baseline mixed-case single-domain grapheme-based RNN-T (Model C).

Model Index	**Vocab**	**Training Data**	**UPWR**	**UPSR**
A	Grph	lowercase	0.11	0.48
B	WPM	lowercase	0.13	0.43
C	Grph	mixed-case	0.16	0.68
D	WPM	mixed-case	0.21	0.82
Cnorm	Model C + lowercaseFST		0.11 0% wrt A	0.47 -2.08% wrt A
Dnorm	Model D + lowercaseFST		0.13 0% wrt B	0.56 +30.2% wrt B

Table 3: *Stability of WPM and grapheme RNN-T models with training data from a single domain. "Grph" is abbreviation for grapheme.*

ID	Vocab	PEI	Training Domain & Data	WER	PER	mWER	UPWR	UPSR
C	Grapheme	50	Single	4.7(0%)	2.5	6.9(0%)	0.16	0.68
D	WPM	50	Single	4.3(-8.51%)	2.8	6.6(-4.3%)	0.21	0.82
E	WPM	50	Multi	4.0(-14.9%)	3.1	6.7(-2.9%)	0.37	1.19
F	WPM	50	Multi + Text Norm	3.8(-19.1%)	2.6	6.2(-10.1%)	0.26	0.96
G	WPM	50	Multi + Text Norm + Domain-id	**3.6(-23.4%)**	**2.3**	**5.8(-15.9%)**	0.14	0.48
H	WPM	200	Multi + Text Norm + Domain-id	**3.6(-23.4%)**	**2.3**	**5.8(-15.9%)**	**0.04**	**0.15**

Table 4: *RNN-T model quality and stability. PEI: partial emission interval (ms); PER: punctuation error rate; mWER: mixed-case WER.*

5. Improving Streaming Stability

5.1. Stability vs Partial Emission Intervals

Previous works have shown trade-offs between partial emission intervals (PEI) and model stability [3, 4, 5, 2]. PEI is the time we set between showing consequtive partial results from the ASR recognizer. Naturally, the longer a partial lives in the lattice beam, the less likely the partial word is going to be revised.

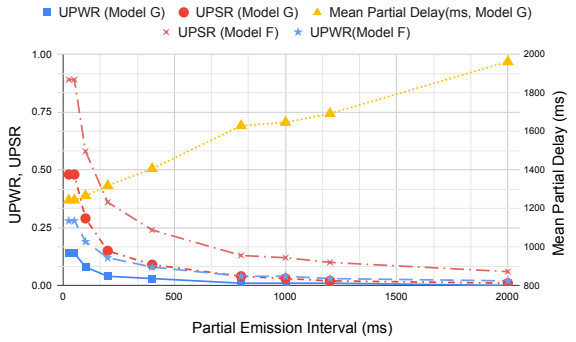

Figure 1: *Instability (UPWR, UPSR) with respect to partial emission intervals (ms) in Model G and Model F. Mean partial delay (ms) of Model G is also shown.*

We thus explore the relationship between partial emission intervals (in ms) and model stability with a streaming RNN-T E2E speech recognizer. Figure 1 shows how the instability metrics decrease logarithmically when the partial emission interval lengthens. In Section 4, we introduce methods to stablize WPM target RNN-T models by pushing the UPWR and UPSR curves to a lower magnitude. For instance, by using domain-ID, we show improvement of stability from model F to model G at no delay of the partial emission intervals at every point along the UPWR and UPSR curves in Figure 1. In this section, we show how stability can be improved by sliding to different PEI's along the curves. We also measure the latency of the speech recognizer in terms of mean partial delay, which is the average time before each hypothesized word is first shown on the screen after the beginning of the utterance.

Changing the partial emission interval from 50ms to 200ms results in 75ms increase in mean partial delay (see Figure 1). However, at 200ms PEI, as shown in Table 4, model H has 71.6% improvement in UPWR, and 68.8% improvement in UPSR, compared to model G, achieving a good tradeoff between latency and stability.

5.2. Stability Thresholding vs Mean Partial Delay Tradeoff

In [5], McGraw and Gruenstein developed a logistic regression approach to estimate the stability of partial results. We analyze the impact on model stability by thresholding the stability scores of partials. Each partial is scored based on the proposed logistic regression model, and partial words that have stability score exceeding the threshold are immediately shown on the screen while the others are withheld. Figure 2 shows that increasing the threshold improves model stability. UPWR and UPSR drop sharply when the threshold grows from 0.1 to 0.2, and decrease linearly when the threshold score is bigger than 0.2.

Accompanying the precipitous drop in the speech recognizer's UPSR and UPWR is a perceptible increase in recognizer latency. This is measured by a 18.6% increase of the mean partial delay from 1243ms to 1474ms. This delay is expected because the logistic regression model depends on a feature, **age**, which measures the length of time that a partial survives the best decoding path. Partials at the beginning of an utterance usually do not have long **age**, and thus are predicted to be less stable.

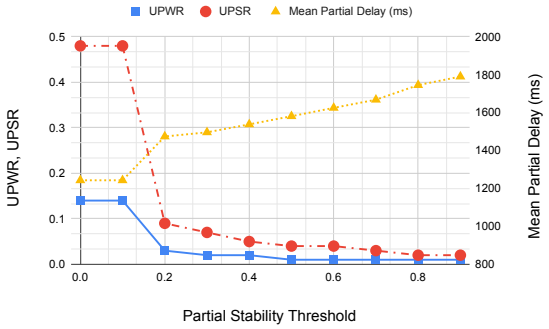

Figure 2: *UPWR, UPSR and mean partial delay with respect to the threshold of stability score in Model G.*

6. Conclusion

In this paper, we describe the problem of instability in the framework of a streaming E2E RNN-T based multi-domain ASR. We observe that model training methods such as mixing data from multiple speech domains and adopting WPM-targets have adversely impacted on the stability of the streaming RNN-T models. In addition, the introduction of text normalization features, including mixed-case data, numerics and spoken punctuations, has also resulted in model stability degradation. We first introduce a novel set of metrics, UPWR and UPSR, to quantify the magnitude and frequency of instability occurences. We then categorize instability into 5 types. We outline model training techniques that dramatically improve the stability of streaming E2E models at no delay to the output words. We show partial emission delay is an effective tool to reduce streaming instability of the RNN-T models but only up to a point before the partial emissions delays become perceptible.

7. Acknowledgements

We thank Arun Narayanan and Tara Sainath for their efforts in developing the RNN-T models with multi-domain training data, text normalization of training data, and the domain-id inputs. We also appreciate Dirk Padfield's careful review of our drafts.

8. References

[1] E. O. Selfridge, I. Arizmendi, P. A. Heeman, and J. D. Williams, "Stability and accuracy in incremental speech recognition," in *Proceedings of the SIGDIAL 2011 Conference*. Association for Computational Linguistics, 2011, pp. 110–119.

[2] N. Arivazhagan, C. Cherry, W. Macherey, P. Baljekar, G. Foster *et al.*, "Re-translation strategies for long form, simultaneous, spoken language translation," *arXiv preprint arXiv:1912.03393*, 2019.

[3] G. A. Fink, C. Schillo, F. Kummert, and G. Sagerer, "Incremental speech recognition for multimodal interfaces," in *IECON'98. Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society (Cat. No. 98CH36200)*, vol. 4. IEEE, 1998, pp. 2012–2017.

[4] T. Baumann, M. Atterer, and D. Schlangen, "Assessing and improving the performance of speech recognition for incremental systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 380–388.

[5] I. McGraw and A. Gruenstein, "Estimating word-stability during incremental speech recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] Y. Shangguan, J. Li, L. Qiao, R. Alvarez, and I. McGraw, "Optimizing speech recognition for the edge," *Third Conference on Machine Learning and Systems, On-Device Intelligence Workshop*, 2020.

[7] A. Narayanan, R. Prabhavalkar, C.-C. Chiu, D. Rybach, T. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *IEEE Automatic Speech Recognition and Understanding (ASRU)*, 12 2019, pp. 920–927.

[8] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen, C.-C. Chiu, D. Garcia, A. Gruenstein, K. Hu, M. Jin, A. Kannan, Q. Liang, I. McGraw, C. Peyser, R. Prabhavalkar, G. Pundak, D. Rybach, Y. Shangguan, Y. Sheth, T. Strohman, M. Visontai, Y. Wu, Y. Zhang, and D. Zhao, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *International Conference on Acoustics, Speech, and Signal Processing (ICCASP 2020)*, 2020.

[9] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 441–447.

[10] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[11] X. Chen and B. Xu, "Stable-time prediction during incremental speech recognition," in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*. IEEE, 2016, pp. 135–138.

[12] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 193–199.

[13] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5149–5152.

[14] M. Chua, D. Van 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)*, 2018.