# Adjunct-Emeritus Distillation for Semi-Supervised Language Model Adaptation

*Scott Novotney*[*], *Yile Gu*[*], *Ivan Bulyko*

Amazon Alexa, USA

snovotne@amazon.com, yilegu@amazon.com, ibbulyko@amazon.com

## Abstract

To improve customer privacy, commercial speech applications are reducing human transcription of customer data. This has a negative impact on language model training due to a smaller amount of in-domain transcripts. Prior work demonstrated that training on automated transcripts alone provides modest gains due to reinforcement of recognition errors. We consider a new condition, where a model trained on historical human transcripts, but not the transcripts themselves, are available to us. To overcome temporal drift in vocabulary and topics, we propose a novel extension of knowledge distillation, *adjunct-emeritus distillation* where two imperfect teachers jointly train a student model. We conduct experiments on an English voice assistant domain and simulate a one year gap in human transcription. Unlike fine-tuning, our approach is architecture agnostic and achieves a $14\%$ relative reduction in perplexity over the baseline approach of freezing model development and improves over the baseline of knowledge distillation.

**Index Terms**: language modeling, automatic speech recognition, semi-supervised modeling

## 1. Introduction

To improve customer privacy, commercial speech recognition systems are reducing human transcription and relying solely on automatically transcribed data to train acoustic and language models. In this work, we propose adaptation methods that utilize models trained on historical data, while making that data unavailable for new language model training or for re-training the teacher model. Under this scenario, we assume that one can access the parameters and likelihoods from the historically trained model, but not the underlying data. We consider model inversion and model attacks to be outside the scope of this work.

In our experiments, we simulate a scenario where one year has passed since the teacher model was trained using historical transcripts, and no new human transcriptions had taken place. We have access to some recent transcriptions in the form of a development set (for LM weight tuning) and a test set (for reporting performance). Available to us is 1) a language model trained on historical transcripts before they were deleted; 2) a large set of machine transcribed audio recent in time to the test set; and 3) a small manually transcribed tuning data set. We do not consider acoustic modeling or all neural systems in this work, though they are interesting extensions of this approach. Prior work has shown that language models are more sensitive to machine transcription errors than acoustic models [1]. Unlike most prior work, we take into account that human transcription and test data are temporally evolving for voice assistant applications and consider the scenario where our test data is collected one year after human transcription ceased.

[*]equal contribution

With only a historical language model available, we cannot use methods such as data concatenation or pre-training on machine transcripts and then fine-tuning on accurate human transcripts. Additionally, due to temporal drift of entities and topics, ceasing to update the LM will result in poor performance on recent test data. Relying solely on machine transcripts is also sub-optimal due to reinforcement of recognition errors.

We propose a novel extension of knowledge distillation [2] that learns from two teachers. The "adjunct": trained from recent, but machine generated (and therefore erroneous) transcripts. The "emeritus": trained from human generated, but out of date, transcripts. These two teachers jointly train a single student architecture. Through the use of word piece segmentation and knowledge distillation, our approach is model agnostic, allowing for improved architectures in the student model while still benefiting from the emeritus model. In contrast to adaptation methods like fine-tuning, we are not tied to the architectures of historically trained models. Our contributions are 1) quantify the trade-offs of increasing customer privacy and impact on language modeling and 2) a novel semi-supervised adaptation technique for producing a student model without requiring historical transcripts.

## 2. Prior Work

Knowledge distillation [2] is a method of model compression to distill a large model or ensemble of models into a single, faster, architecture. Extending from [3], Hinton et al. distill an ensemble of ten deep neural networks for acoustic modeling into one single architecture without loss of WER. Knowledge distillation for an acoustic model [4] resulted in an 11% relative word error rate reduction (WERR) using a larger and more accurate teacher model across four different languages. A 3x slower decoder resulted in a 10% WERR reduction and doubling the machine transcripts gave a 3% improvement. Similar efforts in end-to-end ASR self-training [5] reported that equal weighting of supervised and machine transcribed corpora can achieve up to 5% WERR when ensembling six models. Our approach uses knowledge distillation to access the learned knowledge of data that is no longer accessible, instead of model compression.

The most similar approach to our method is semi-supervised or self-training of speech recognition models. Earlier efforts at back-off language model adaptation successfully incorporated automatically decoded transcripts into n-gram language models [6, 7], but with modest WER gains. Initial work with HMM acoustic models and n-gram LMs showed much larger gains for acoustic rather than language modeling [8]. Efforts at semi-supervised language modeling demonstrated a 3.5% WERR when adapting a backoff LM using lattice-based fractional counts [9]. Contrasting between AM and LM self-training, recent work [10] concluded that active learning helps the AM, but did not improve the LM.

While the sub-task of *unsupervised* modeling bootstraps ASR without any manual or in-domain transcripts [11], we do not consider the task of a new domain, but of continuing strong ASR recognition while respecting customer privacy. Adversarial attacks for speech recognition [12] demonstrate how models can "leak" sensitive information about customers, but we consider this outside the scope of our work.

The field of life-long machine learning has produced ideas for future work [13] to train one architecture capable of multiple NLP tasks. A "memory bank" of previously seen examples is retained then interleaved at adaptation time to ensure that the one model can accurately perform shifting NLP tasks. Their approach outperforms on-line adaptation across a variety of NLP tasks.

## 3. Technical Approach

Our objective is to estimate the parameters, $\Theta$, of a neural language model so that we may compute $P_\Theta(w_1, w_2, \ldots, w_N)$ for word sequence $w_1$ to $w_N$. Available to us is a model trained on historical transcripts and recent adaptation data (a large volume of machine transcribed text in this work). Since we wish to produce one architecture, we rule out approaches such as system interpolation at inference time. Likewise, we cannot employ the typical strategy of learning from machine transcription by first pre-training our model or embedding layers since we do not have the historical transcripts available to fine-tune the model. We first describe knowledge distillation and then introduce our extension of this approach.

### 3.1. Knowledge Distillation

Knowledge distillation [2] compresses information of a *teacher* model to a (typically more efficient) *student* model by using the teacher's posteriors as training targets. Given training sequence $w_1, w_2, \ldots, w_N$, we generate logits from the teacher model $z_{t,1}, z_{t,2}, \ldots, z_{t,N}$ through a forward pass of the teacher model. To better capture the model's alternate hypotheses, the model posteriors, $P_t$, are smoothed with a constant temperature $T$ as

$$P_t(w) = \frac{\exp(z_t(w))/T}{\sum_{w'} \exp(z_t(w')/T)}. \quad (1)$$

The student parameters $\Theta$ are optimized through stochatic gradient descent to minimize a weighted combination of the Kullback-Leibler divergence between the model's current posteriors, $P_\Theta$, and the teacher posteriors, $P_t$. Additionally, the KL divergence term between the empirical distribution of the target sequence, $\tilde{P}$ is added since prior work demonstrated empirically that it is a useful regularizer[1], with $\alpha$ set as a hyper-parameter,

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \left[ \mathrm{KL}(P_{\Theta,i} || P_{t,i}) + \alpha \cdot \mathrm{KL}(P_{\Theta,i} || \tilde{P}_i) \right], \quad (2)$$

### 3.2. Adjunct Emeritus Distillation

Our approach now contains two teachers: the *emeritus* teacher: an LM trained on historical manual transcripts whose parameters are frozen; and the *adjunct* teacher: an LM trained on recent machine transcripts. We jointly distill their logits to the student model through an additional term in Eqn. 2. We first compute the forward pass of the training sequence through the emeritus and adjunct models and compute the smoothed posteriors, $P_{\text{em}}$ and $P_{\text{adj}}$, using temperature as above. Then, we compose a joint loss function for the parameters of the student model, $\Theta$,

$$\mathcal{L}(\Theta) = \mathrm{KL}(P_\Theta || P_{\text{em}}) + \alpha \mathrm{KL}(P_\Theta || P_{\text{adj}}) + \beta \mathrm{KL}(P_\Theta || \tilde{P}) \quad (3)$$

where $\alpha$ and $\beta$ are hyper-parameters. This approach lets the scientist intuitively set $\alpha$ as the relative value of the adjunct to the emeritus model. When $\alpha = 0$, this is the same as knowledge distillation.

We used a temperature of 1.01 for both the emeritus and adjunct model and manually evaluated different hyper-parameter weights, with $\alpha = 0.05$ and $\beta = 2$ being the most optimal. We attempted to automatically adjust $\alpha$ (the relative importance of the adjunct model to the emeritus model) based on signals from the two models. A third logistic regression model (the model combiner) was trained on entropy of each of the models' posteriors as well as the final hidden layer after a forward pass. This model was used to adapt the $\alpha$ per target. Unfortunately, we saw no benefit for this approach when we trained the third system combination model on machine transcripts. If we had manual data available for tuning the model combination, there was a substantial gain, but this was not a fair experiment given the constraints of our scenario. Future work could explore how to train an adaptive logistic regression approach on historical data.

All our language models used a two layer uni-directional LSTM [14] with 512 dimensional embedding and hidden units. We initialized all parameters randomly from $\mathcal{N}(0, 1)$ and hidden states to 0. Models were trained using Adam [15] with an initial learning rate of 0.01. The learning rate was cut by a factor of 2 each time loss plateaued on dev data, using early stopping to cease training. Although our distillation approach is architecture agnostic, we found no benefit on our test set for transformer models or larger/deeper LSTM architectures. We use 10,000 unit word pieces trained on human transcripts [16] for our vocabulary. This allows for logits to be comparable across model architectures while ensuring no out of vocabulary words.

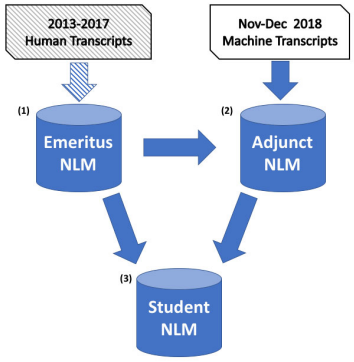


Figure 1: Overview of Adjunct-Emeritus approach. *The* Emeritus *(1) model is trained on historical data that is no longer accessible. The* Adjunct *(2) model is a fine-tuned version of the Emeritus on recent high-confidence machine transcripts. The smoothed logits from both architectures are used to train the* Student *(3) model on sequences from the recent machine transcripts. Section 5 contrasts this approach with alternate adaptation methods and will show that this is our best performing system.*

[1] this is the same as usual cross entropy loss, but we use KL divergence for notation consistency.

# 4. Data and experimental setup

Table 1: *Corpora used in this work. We simulate a one year gap between ceasing transcription (row 1), adapting with recent data (rows 2/3) and measure WER on test data (row 4).*

| Name | Date Range | Label | Hours |
|---|---|---|---|
| Historical | Up to 2017 | Human | 19K |
| Adapt-MT | Fall 2018 | Machine | 230K |
| Adapt-HT | Fall 2018 | Human | 5K |
| Tune | Jan '19 | Human | 276 |
| Test | Jan '19 | Human | 92 |

We use a corpus of non-identifiable voice assistant recordings for experimentation. To accentuate the effects of temporal drifts of topics and entities, we simulated the scenario where human transcriptions are no longer available. Specifically, we simulated transcription ceasing in December 2017; machine transcripts from Fall 2018 were available; and we evaluated the model's performance on manual transcripts from January of 2019. We assumed a small amount of development data was available for tuning and determining early stopping for learning rate convergence.

The machine transcripts were selected to have recognition confidence estimated to be below average WER. We re-sampled the data to reflect the overall distribution of topics from the entire set of audio recordings. The set of topics came from a natural language understanding (NLU) tagger that categorized recordings into 21 "intents" such as Music, Information, and None. This re-sampling controls for short, high confidence, recordings dominating the machine transcript data set. As a contrasting experiment, we also used a set of human transcribed data covering the same time range as the machine transcription adaptation data. Table 1 describes the data used.

We decoded the test set using a hybrid ASR model to generate 10-best hypotheses for our neural LMs to rescore. In the first-pass, the acoustic model was a low-frame-rate model with 2-layer frequency LSTM [19] followed by a 5-layer time LSTM trained on cross-entropy loss, followed by sMBR loss [20]; the first pass language model used was a Kneser-Ney (KN) [18] smoothed n-gram language model trained on a large variety of in-domain and out-of-domain corpora. We then replaced the LM scores with our neural LMs and combined with AM scores to perform second pass re-scoring.

Section 5 details experiments of multiple models, we refer to them as below. Methods using knowledge and adjunct-emeritus distillation produce a student model.

- *emeritus*: Trained on historical transcripts (2013 to 2017).
- *KD*: Vanilla knowledge distillation of *emeritus* using adapt-MT corpus.
- *adjunct*: Trained on machine transcripts from adapt-MT.
- *adjem*: Our approach of adjunct-emeritus distillation combining *emeritus* and *adjunct* using adapt-MT.
- *adjunct-ft*: Start with *emeritus* model, then fine-tuned on adapt-MT.
- *adjem-ft*: Our approach, using *adjunct-ft* model and *emeritus*, adapted with adapt-MT.
- *concat-mt*: An upper bound model trained on historical and adapt-MT data concatenated together.
- *concat-ht*: An upper bound model trained on historical and adapt-HT data concatenated together.

# 5. Results and discussions

We evaluated perplexity (PPL) and WER on the Jan 2019 test data (row 5 of Table 1) and report relative perplexity reduction (PPLR) and WER reduction (WERR). All models have the same vocabulary of 10,000 subword units. Results are broken down by "tail" and "head" recordings. *Head* are recordings whose transcripts occur in the top $1\%$ by frequency in the test set after excluding wake words. *Tail* are recordings with transcripts that occur in the test data once. Although we report relative reductions, absolute WERs were less than 10% and Perplexity (PPL) less than 30.

## 5.1. Baseline Models

First, we performed a controlled study to quantify the relative benefit of human versus machine transcripts. We trained 2x512 LSTMs on each of the first three data sets in Table 1. We also trained a fourth model that replaced human transcripts with machine transcripts in the set of recordings – the volume of data was the same, but transcripts were machine generated. Table 2 details the results.

Table 2: *Comparison (in relative change) between adjunct models (Adapt-HT, Adapt-MT, Adapt-MT*) trained on human transcripts and MT for the same time period. Row 4, Adapt-MT*, uses the same recordings as row 2, Adapt-HT, but replaces human transcripts with machine transcripts. Emeritus model trained on historical data is used as the baseline for WERR. Negative WERR indicates degradation. Oracle n-best rescoring WERR is 28%.*

| Model | PPLR | Tail PPLR | Head PPLR | WERR |
|---|---|---|---|---|
| Historical | – | – | – | – |
| Adapt-HT | 4.4% | -7.9% | -2.8% | -0.4% |
| Adapt-MT | -26.3% | -56.4% | 10.0% | -1.7% |
| Adapt-MT* | -42.5% | -46.5% | -32.8% | -3.5 % |

We see that none of the models trained on recent data can match the historical model, due to the large amount of human transcriptions used to train it. Furthermore, a large volume of machine transcripts is still worse than human transcription of recent data. This is due to the large increase in tail perplexity (rare recordings), reinforcing the conventional wisdom that machine generated transcripts have difficulty with accurate recognition in the long tail. Finally, increasing the volume of machine transcripts does overcome the worse recognition quality. Scaling machine transcripts from 19K hours to 230K hours of recordings, the models improve by 1.8% relative.

## 5.2. Adaptation Results

We now explore adaptation approaches to combine the historical model and the adaptation data generated with machine transcripts. Table 3 reports improvements for adaptation methods over these starting models. Our baseline is the emeritus model trained on historical data as it had the lowest WER on recent test data. Our upper bound is to retain access to both historical human transcripts and continued human transcription. We construct a model by concatenating the historical and adapt-HT datasets together and train a neural language model. Ideally, our adaptation will *recover* this improvement while not requiring any human transcription.. We report *WER Recovery* as the fraction of the total possible gain a method is able to recover from baseline to this upper bound. The upper bound of keeping his-

Table 3: *PPL and WER Recovery for adaptation methods defined in Section 4. PPL and WER Recovery is the fraction of the gap between our baseline (row 1) and upper bound oracle of continuing manual transcription (row 8). Note for* Head PPL Rec.*, we use* concat-mt *as our upper bound since* concat-ht *has worse performance than* emeritus*. Our baseline adaptation approach (row 2) is knowledge distillation using the adapt-MT data. Training on machine transcripts (row 3) and jointly distilling a student model (row 4) does not provide any gain in WER despite improvements in recognizing frequent recordings. First fine-tuning the historical model (row 5) and then applying adjunct emeritus distillation to this model (row 5) provides our best result with gains in perplexity.*

| Model | Ovrl WER Rec. | Tail WER Rec. | Head WER Rec. | Ovrl PPL Rec. | Tail PPL Rec. | Head PPL Rec. |
|---|---|---|---|---|---|---|
| *emeritus* | – | – | – | – | – | – |
| *KD* | 20% | -17% | 160% | 19% | 23% | 10 % |
| *adjunct* | -93% | -138% | **260%** | -165% | -270 % | 70 % |
| *adjem* | 0% | -46% | **260%** | 54% | 10 % | 90% |
| *adjunct-ft* | -87% | -125% | 140% | -41% | -104% | 90% |
| *adjem-ft* | **27%** | **8%** | 220% | 92% | 71% | 90% |
| *concat-mt* | 87% | 83% | 240% | 127% | 105% | 100% |
| *concat-ht* | 100% | 100% | 100% | 100% | 100% | -10% |

torical transcripts, but no longer transcribing, *concat-mt*, recovers 87% of improvement of WER, and even surpasses *concat-ht* on Head WER. However, poorer performance in the tail indicates that machine transcription is still not able to replace human transcription due to lower accuracy of machine transcription in the tail.

We report results for five different adaptation methods in rows two through six. Our default adaptation approach is to use knowledge distillation (KD) of the emeritus model using adapt-MT data (machine transcripts). This results in a 20% recovery of the possible upper bound, but a degradation in tail WER of -17%. As reported in Table 2, the *adjunct* model trained just on adapt-MT data is worse across the board. When jointly teaching the student model (model *adjem*) from the *emeritus* and *adjunct* results, we do not see an improvement in overall WER despite strong gains in head WER. This is likely due to over representation of frequent recordings in the confidence filtered machine transcription dataset, despite our efforts to re-sample the data to reflect overall usage by topic.

Starting from the emeritus model and fine-tuning on machine transcription data was better than training on machine transcripts alone (compare *adjunct* to *adjunct-ft*). We then used this fine-tuned model as our adjunct model to combine with the *emeritus*, to jointly train the *adjem-ft* student model. This was our best performing result that not only maintained the improvement on Head WER, but also improvd Tail WER, resulting in a 27% overall WER Recovery, the highest WER Recovery among these five methods considered. This model yielded the lowest PPL in overall as well as Tail and Head conditions. By adapting the emeritus model to recent data, the resulting model better captured recent shifts in distributions. But because it was trained on errorful transcripts, it was still useful to jointly combine the two logits through adjunct-emeritus distillation. This method outperformed three semi-supervised adaptation strategies: fine-tuning, knowledge distillation, and re-training on machine transcripts alone. Importantly, it produced a model architecture that is independent of the original emeritus model and does not require multiple models in decoding (as would be required in system combination).

## 6. Conclusion

This paper explores a new research area in privacy preserving speech recognition that aims to find a middle ground between deleting all data and relying solely on machine transcription versus keeping human transcripts forever. Our novel approach,

Table 4: *Example song names corrected by adapting with machine transcription.* **bold** *highlights correct word and* italic *is a mis-recognition by the emeritus model.*

| Release Date | recording |
|---|---|
| July 2018 | play **keke**/*kiki* do you love me |
| Aug 2018 | play **dame tu**/*dummy* to cosita |
| Sept 2018 | play **taki taki**/*top party* |
| Sept 2018 | play **shallow**/*shadow* from a star is born |

Emeritus-Adjunct distillation, allows old historical transcripts to be deleted and still improve neural language models on recent machine transcribed speech data. This improves customer privacy while also benefiting from historical human transcripts and learning from recent machine transcripts. Compared to other approaches adaptation approaches like fine-tuning or vanilla knowledge distillation, our approach does better on the long-tail words, a particular weakness of learning from machine produced transcripts.

We analyzed recordings that most improved from incorporating recent machine transcriptions. The broadest category was better recognition of recent song names. For instance, all these examples wtere songs released in July 2018 or later – six months after our historical data ceased transcription.

In Section 3, we described an adaptive interpolation weight during knowledge distillation that put more weight on the adjunct or emeritus model depending on observed features. Unfortunately, we were only able to see a gain when this third model was tuned on human transcribe data, violating the assumptions of our scenario. Future work can explore approaches to, in a privacy preserving method, train an adaptive weighting based on features such as posterior entropy, attention over prior context, and contextual features like time of day. For example, training a model on heldout historical data or on synthetic data sampled from the historical language model. Section 5 demonstrated that concatenating historical transcripts and recent automatic transcripts results in gains for the tail, middle and head of recordings. However, training one model on the historical transcripts results in a net loss on the tail. This gap of 5% WERR is an opportunity for future work to explore better privacy preserving information in historical transcripts. One potential application is episodic memory [13] to better store data in a granular representation that is not human interpretable. Similarly, applications of privacy preserving machine learning could provide provable representations of transcripts that are not susceptible to adversarial attacks.

# 7. References


[1] Sree Hari Krishnan Parthasarathi and Nikko Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6670–6674.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[4] Olga Kapralova, John Alex, Eugene Weinstein, Pedro Moreno, and Olivier Siohan, "A big data approach to acoustic model training corpus selection," 2014.

[5] Jacob Kahn, Ann Lee, and Awni Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7084–7088.

[6] Michiel Bacchiani and Brian Roark, "Unsupervised language model adaptation," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. IEEE, 2003, vol. 1, pp. I–I.

[7] Gokhan Tur and Andreas Stolcke, "Unsupervised languagemodel adaptation for meeting recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–173.

[8] Scott Novotney, Richard Schwartz, and Jeff Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4297–4300.

[9] Vitaly Kuznetsov, Hank Liao, Mehryar Mohri, Michael Riley, and Brian Roark, "Learning n-gram language models from uncertain data," 2016.

[10] Thomas Drugman, Janne Pylkkonen, and Reinhard Kneser, "Active and semi-supervised learning in asr: Benefits on the acoustic and language models," *arXiv preprint arXiv:1903.02852*, 2019.

[11] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 398–403.

[12] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5231–5240.

[13] Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama, "Episodic memory in lifelong language learning," *arXiv preprint arXiv:1906.01076*, 2019.

[14] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[15] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] Taku Kudo and John Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[17] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister, "Improving asr confidence scores for alexa using acoustic and hypothesis embeddings.," 2019.

[18] Reinhard Kneser and Hermann Ney, "Improved backing-off for m-gram language modeling," in *1995 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1995, vol. 1, pp. 181–184.

[19] Bo Li, Tara N Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean K Chin, et al., "Acoustic modeling for google home.," in *Interspeech*, 2017, pp. 399–403.

[20] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, 2013, vol. 2013, pp. 2345–2349.