Ensembles of Hybrid and End-to-End Speech Recognition

Aditya Kamlesh Parikh, Louis ten Bosch, Henk van den Heuvel
Center for Language and Speech Technology (CLST)
Radboud University, Nijmegen, The Netherlands
{aditya.parikh, louis.tenbosch, henk.vandenheuvel}@ru.nl

Abstract

We propose a method to combine the hybrid Kaldi-based Automatic Speech Recognition (ASR) system with the end-to-end wav2vec 2.0 XLS-R ASR using confidence measures. Our research is focused on the low-resource Irish language. Given the limited available open-source resources, neither the standalone hybrid ASR nor the end-to-end ASR system can achieve optimal performance. By applying the Recognizer Output Voting Error Reduction (ROVER) technique, we illustrate how ensemble learning could facilitate mutual error correction between both ASR systems. This paper outlines the strategies for merging the hybrid Kaldi ASR model and the end-to-end XLS-R model with the help of confidence scores. Although contemporary state-of-the-art end-to-end ASR models face challenges related to prediction overconfidence, we utilize Renyi's entropy-based confidence approach, tuned with temperature scaling, to align it with the Kaldi ASR confidence. Although there was no significant difference in the Word Error Rate (WER) between the hybrid and end-to-end ASR, we could achieve a notable reduction in WER after ensembling through ROVER. This resulted in an almost 14% Word Error Rate Reduction (WERR) on our primary test set and an approximately 20% WERR on other noisy and imbalanced test data.

Keywords: hybrid ASR, end-to-end ASR, confidence measure, entropy, ROVER

1. Introduction

In machine learning, combining multiple models is a common technique used to improve accuracy, robustness, and overall performance of predictive models – this approach is known as ensemble learning. There are several ensemble techniques like bagging (Opitz and Maclin, 1997), boosting (Schwenk and Bengio, 2000), stacking (Pavlyshenko, 2018) and voting (Fiscus, 1997) each with its own advantages and use cases. System combination is a widespread practice in ASR. The fused ensemble system typically yields lower WER compared to individual systems and demonstrates improved generalization to unseen data.

In ASR, the ROVER technique minimizes WER by harnessing distinct error patterns from multiple ASR systems (Fiscus, 1997). The ROVER algorithm first aligns the hypotheses and then chooses the hypotheses based on weighted word level confidence scores. Such ROVER ensemble techniques in ASR have been applied in various contexts, such as language identification (LID), dialectal and accented ASR, and commercial ASRs. (Metze et al., 2000) demonstrates that confidence-based LID outperformed traditional score-based methods, employing confidence scores from multiple monolingual Hidden Markov Model (HMM)-based ASR models. Moreover, hybrid Kaldi ASR models have been combined with ROVER to enhance ASR performance and robustness in (Audhkhasi et al., 2013; Jalalvand et al., 2015; Gebauer et al., 2023; Yamini and Ingo, 2021). Valente (2010) explored the use of the Dempster–Shafer (DS) combination rule for multi-stream ASR, boosting recognition accuracy and resilience in adverse acoustic conditions.

Since 2021, end-to-end architecture-based ASR systems gained traction with substantial success in terms of WER. However, the computation of word-level confidence in end-to-end Connectionist Temporal Classification (CTC)(Graves et al., 2006) based models is a challenge compared to the computation in hybrid ASR systems. End-to-end ASR systems determine confidence in predictions by analyzing softmax probabilities of output vocabulary units (logits), with the highest probability indicating confidence (Hendrycks and Gimpel, 2017). However, this approach is problematic due to "prediction overconfidence", in which the probability distribution heavily favors the most supported hypothesis (which may be incorrect). In the case of CTC based end-to-end ASR, this "prediction overconfidence" is observed when an *incorrect* prediction is assigned a probability higher than 0.9 (Laptev and Ginsburg, 2023). Another issue is the *prediction granularity*, where speech applications typically demand confidence assessments at the word level, while end-to-end ASR systems produce outputs per time frame. This granularity issue is why finding an *aggregation* method for lifting, e.g., frame or grapheme level scores to word-level scores is important as well.

Valente (2010) describes the use of inverse-entropy to combine multiple multi layer perceptron (MLP) classifiers trained on different representations of the speech signal, and recently Laptev

LREC-COLING 2024, pages 6199–6205
20-25 May, 2024.

and Ginsburg (2023) employed advanced entropy-based methods to measure confidence scores and reduce the risk of overconfidence in conformer CTC (Gulati et al., 2020) and recurrent neural network transducer (RNN-T) models. A recent study by Gitman et al. (2023) used confidence measures based on entropy to combine multiple conformer CTC and RNN-T models. As far as we know, there is no study where entropy-based confidence scores are used to combine HMM-DMM based hybrid ASR models with end-to-end ASR models. In this paper, we try to narrow this gap by investigating an ensemble model by combining CTC based wav2vec 2.0 (Baevski et al., 2020) XLS-R (Babu et al., 2022) model with HMM-DNN Kaldi (Povey et al., 2011) using ROVER.

2. Data

2.1. Audio Data

We employed the same audio data set for training the Kaldi-based hybrid ASR models and for fine-tuning the CTC-based end-to-end wav2vec 2.0 XLS-R model. To that end we combined three small open-source Irish datasets: (1) the Common Voice (CV) Irish dataset (Ardila et al., 2020) served as primary source. To ensure data quality, we exclusively utilized validated utterances from CV dataset, excluding those earmarked for the test set. (2) The "Living Audio" (LA) dataset (Braude et al., 2019), which contributed an additional hour of Irish speech data. (3) All Irish utterances available from the "Google Fleurs" (GF) dataset (Conneau et al., 2023). By combining these three datasets, we compiled a audio training dataset comprising 9,274 utterances (13.5 h).

For testing purposes, we used two sets: (1) the CV Irish 'Test' set, containing 513 utterances (0.5 hours of speech), and (2) the 'Invalidated' CV Irish utterances, encompassing 282 utterances (0.3 hours of speech) after filtering out samples with very high background noise or without any speech. The 'Invalidated' clips in the CV Irish dataset are those with more downvotes than upvotes, implying they may contain significant background noise, incorrect utterances spoken compared to the original transcript, resulting in a higher WER compared to the CV Test set. The reference transcripts in train and test set consist of total 33 tokens, including 18 from the Irish original alphabet, 6 from the English alphabet for foreign words, and 5 accented vowels apart from that, there is <pad>, <unk>, aphostrophe and a <blank> token. An overview of the Irish speech data sets is provided in Table 1.

Dataset	#Utterances	Duration	#Word Tokens	#Word Types
CV Train	4097	4.1h	27880	2341
LA Irish	1122	1h	11360	3542
GF Irish	1947	8.4h	48929	9866
CV Test	513	0.5h	3423	1109
CV Invalidated	282	0.3h	2230	707

Table 1: *Overview of the Irish datasets used. The abbreviations **CV**, **LA** and **GF** denote Common Voice, Living Audio and Google Fleurs, respectively.*

2.2. Text Data and Pronunciation Lexicons

For the language model (LM), we used the CC-100 Monolingual datasets sourced from Web Crawl Data (Conneau et al., 2020). This extensive resource covers over 100 languages, including Irish, and comprises a total of 84 million word tokens and 0.12 million word types, each with a frequency higher than 10. For our experiments with Kaldi ASR, we trained a Grapheme-to-Phoneme (G2P) model based on Joint-sequence models, by using 13,300 seed Irish pronunciations extracted from Wikipron (Lee et al., 2020).

3. Methodology

To train the (lattice based) hybrid ASR we used Kaldi ASR toolkit's mini-librispeech recipe[1]. The acoustic model (AM) is a combination of a Time-Delayed Neural Network (TDNN) and a Convolutional Neural Network (CNN). Next, a 4-gram statistical LM for Irish was generated using the SRILM tool (Stolcke, 2002), based on the text resources mentioned in section 2.2. Finally, the pronunciation lexicons were created using a data-driven approach for Irish as explain in the section 2.2. For decoding with Kaldi we used Minimum Bayes Risk (MBR) decoding to select the most likely candidate hypothesis with the lowest expected loss under a probability model. It does so by minimizing the expected classification error, effectively incorporating the loss function into the decision-making process. MBR decoding is typically implemented by re-ranking a list of N-best transcriptions generated by an initial decoder. In order to do so we used the `lattice-to-ctm-conf` script[2] with the `--decode-mbr` flag set to `true`. The MBR method derives the most probable transcript w^* by optimizing a function of the following form:

$$w^* = \arg\min_{w} \sum_{w'} p(w'|x)L(w, w')$$

[1] github.com/kaldi-asr/kaldi/blob/master/egs/mini_librispeech/

[2] github.com/kaldi-asr/kaldi/blob/master/src/latbin/lattice-to-ctm-conf.cc

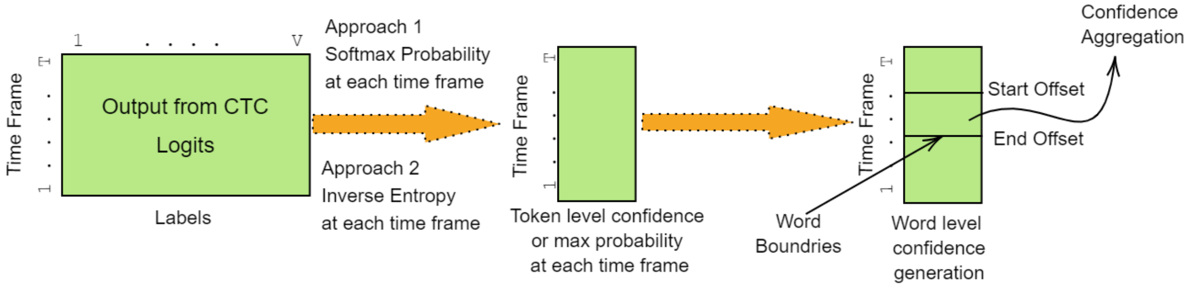

Figure 1: *Representation of the two approaches for estimating confidence scores for CTC based end-to-end ASR. From CTC output, we obtain the probability distribution of each token vocabulary for each time-frame. Based on mentioned two approaches we extract confidence score at token level and finally aggregate to obtain scores at word level using word boundaries. The size of the token vocabulary is denoted by V, the number of tokens is denoted by T.*

In the above equation, $p(w'|x)$ represents the probability of the word sequence w' given the audio signal x, and $L(w, w')$ denotes the Levenshtein distance between the two word sequences w and w'.

We finetuned an end-to-end CTC based model on the training audio dataset (section 2.1). We utilized the publicly released pre-trained wav2vec 2.0 model, XLS-R, which was trained on 436K hours of publicly available speech audio and is available on HuggingFace[3]. We used the 300 million-parameter version of XLS-R[4], which is among the smaller versions (models range from 300 million to two billion parameters). The fine-tuning was performed on an NVIDIA Tesla T4 GPU using the Adam optimizer, with a learning rate starting with a warm-up for 500 steps, peaked at $3e^{-4}$ for all global steps, and then decayed exponentially. The total number of global steps for fine-tuning to Irish was 7180. We employed greedy decoding, and simply picked up the best hypothesis at each time step.

To compute the confidence scores for the end-to-end CTC based wav2vec 2.0 model, we used two approaches mentioned in the figure 1. In the first approach, we determine the maximum probability for each token and aggregate the word-level probability based on a word's onset and offset. The maximum probability (henceforth noted as max prob) is calculated using the `log_softmax` function with a temperature scaling factor of 1. When the temperature is set to one, scaling is not applied to max prob, and any entropy type behaves like the Shannon entropy. In the second approach, we first converted logits to `log_softmax`, assigned an entropy value to each predicted token (here: grapheme), and used the inverse entropy as a 'confidence' value. We then aggregated the token-level confidence scores to word-level scores based on onsets and offsets of each hypothesized word. In this entropy-based approach, we followed the method described by Laptev and Ginsburg (2023). For each token's probability distribution p_v, vocabulary size V and at time frame t_i, we computed the confidence score $C(t_i)$ based on exponentially normalized Renyi's entropy via:

$$C(t_i) = \frac{(\sum e^{p_v \cdot \tau})^{\frac{1}{\tau - 1}} \cdot V - 1}{V - 1}$$

In this formula, τ denotes a temperature scaling factor. Temperature scaling involves the multiplication of log-softmax values by τ (τ between 0 and 1). We tuned Renyi's entropy based confidence by adjusting τ (Hinton et al., 2015). While this approach is often employed to recalibrate raw prediction probabilities, it doesn't result in a significant improvement in confidence itself (Wang et al., 2020). However, this adjustment makes the resulting confidence score more compatible with other confidence scores such as the ones used in the Kaldi-based (lattice) methods, and thereby enhances the robustness of entropy-based confidence measures in ensemble classification.

To convert token-level confidence to word level confidence score we used three aggregation methods, here denoted $mean$, $minimum$ and $product$ referring to the math operation involved. In CTC ASR models a special blank token `<blank>` is used. In our aggregation approaches, we excluded `<blank>` tokens prior to the aggregation of the token-level confidence scores.

We computed mean and standard deviation (SD) of confidence scores of both Kaldi and the end-to-end ASR. For the end-to-end ASR, we computed token probabilities, and tuned the Renyi's entropy confidence with different aggregation methods.

3.1. ROVER

After computing the confidence scores for both systems, we merged the hypotheses from both ASRs using ROVER (Fiscus, 1997; Yamini and Ingo, 2021). ROVER combines transcriptions from

[3] `huggingface.co/docs/transformers/model_doc/wav2vec2`

[4] `huggingface.co/facebook/wav2vec2-xls-r-300m`

multiple ASR systems by selecting the best word hypothesis for each word through a voting process. This process has two phases. In the first phase, ROVER aligns the hypotheses by minimizing the total cost of word insertion, deletion, and substitution to make all hypotheses identical. Alignments are done iteratively, and if a word is missing in the alignment, it's replaced with a null word transition. In the second phase, it uses a voting mechanism to calculate a word-level score. This score is determined by a weighted sum of the word-level confidence $C(w_i)$ and the number of times the word occurs normalized by total number of hypothesis. In the equation below, word occurrence is represented as $N(wi)$ and total hypothesis is represented as N_s. In our case it is 2 (One from Kaldi and another from wav2vec 2.0). $\alpha \in [0, 1]$ is the weight given to the word occurrence. In our experiments, we gave less weight to word occurrence (0.3) and more to confidence scores (0.7).

$$Score(w_i) = \alpha(\frac{N(w_i)}{N_s}) + (1 - \alpha)C(w_i)$$

The performance of the ASR models and ROVER is expressed in terms of the traditional WER.

4. Results

Table 2 shows confidence scores in various conditions, while Table 3 presents WER scores.

In Table 2, the Kaldi-based confidence scores vary on the Irish Test set with a mean of 0.9435 (SD 0.1479); on the Invalidated test set the mean is 0.8850 (SD 0.2030). Softmax probability of CTC based end-to-end model reflects "over-confidence" by showing mean of approximately 0.999 (SD 0.0132) for both the Irish Test and Invalidated. While using entropy based confidence score in wav2vec 2.0 and tuning it by temperature scale τ, we were able to achieve a mean and SD comparable to Kaldi's confidence with $mean$-based aggregation.

Table 3 shows that Kaldi-based ASR yielded a WER of 26.71% on the Irish Testset and approximately 39% WER on the Irish Invalidated set. When fine-tuned using similar audio data, the wav2vec 2.0 XLS-R model showed a slight enhancement in performance, achieving a WER of 25.81% on the Irish Testset and a roughly 3% reduction in WER on the Irish Invalidated set. Ensembling Kaldi ASR with the softmax probability-based confidence scores from the wav2vec 2.0 XLS-R model did not result in a significant improvement in performance on both test datasets. However, when combining Kaldi's confidence score with the entropy-based $mean$ and $minimum$ aggregated confidence score of the end-to-end wav2vec 2.0 model using ROVER, a significant enhancement was observed, resulting in a substantial 14% WERR achieved on the Irish Testset and approximately 20% WERR on the Irish Invalidated set compared to the Kaldi base model. This improvement also outperformed the fine-tuned wav2vec 2.0 model by 11% and 13% WERR on the Irish Testset and Invalidated set, respectively. Notably, the $product$ aggregation method yielded less improvement in comparison to other two on both test datasets.

Confidence	Test		Invalidated	
	Mean	SD	Mean	SD
Kaldi	0.9435	0.1479	0.8850	0.2030
CTC max prob	0.9990	0.0132	0.9998	0.0141
CTC tuned Renyi's entropy (Exponentially Normalized)				
Mean	0.9329	0.0351	0.8821	0.0480
Min	0.9165	0.0862	0.8586	0.1021
Prod	0.6937	0.1749	0.5304	0.2237

Table 2: *This table compares different confidence methods used for ensemble by ROVER. Mean and SD are provided for different combinations. CTC max prob is the softmax of probability distribution over the labels in the output of wav2vec 2.0. In the CTC tuned Renyi's entropy based confidence, the temperature τ was set to 0.40 and 0.36 for the Test set and Invalidated set, respectively.*

Model	Test	Invalidated
Kaldi base	26.71	38.98
Wav2vec 2.0 base	25.81	35.69
ROVER with max prob	25.60	38.95
ROVER with tuned entropy		
Mean	22.94	31.01
Min	22.97	31.06
Prod	23.21	31.43

Table 3: *Comparing WER for Kaldi ASR, fine-tuned wav2vec 2.0 model, and ROVER on Irish Test and Invalidated sets. The last three rows depict WER improvements through the combination of Kaldi's confidence score and wav2vec 2.0's confidence scores using tuned Renyi's entropy with the three aggregation methods.*

5. Discussion and Conclusion

We present an approach to ensemble the hybrid Kaldi based model with the end-to-end wav2vec 2.0 XLS-R model and demonstrate the significant effect of ROVER in terms of performance for the under-resourced language Irish. This significant improvement can only be achieved by using appropriate methods for combining confidence measures. One major challenge in combining hypotheses lies in calibrating confidence scores by adjustment of the temperature parameter τ. The determination of the optimal value of τ for all datasets is crucial to ensure that the resulting confidence scores can be compared in terms of their mean and SD.

This ensemble approach offers another substantial advantage, namely for mitigating the problems related to the generation of non-lexical words and hallucinations within the end-to-end ASR method. The Kaldi-based hybrid ASR, owing to its restricted lexicon search space based on the pronunciation lexicons employed during training, effectively confines output word prediction. This restriction leads to more precise and customized results, reducing inaccuracies. A drawback of the ensemble method is that it is not suitable in applications where close to real-time responses are crucial. By combining hybrid model with end-to-end model, the computational resource required for decoding will also increase. However we believe, that this approach has the potential to facilitate the development of diverse and specialized ASR systems tailored for specific use cases.

6. Acknowledgements

This work has been conducted in the SignON project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017255.

7. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Kartik Audhkhasi, Andreas M Zavou, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2013. Empirical link between hypothesis diversity and fusion performance in an ensemble of automatic speech recognition systems. In *INTERSPEECH*, pages 3082–3086.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

David A Braude, Matthew P Aylett, Caoimhín Laoide-Kemp, Simone Ashby, Kristen M Scott, Brian Ó Raghallaigh, Anna Braudo, Alex Brouwer, and Adriana Stan. 2019. All together now: The living audio dataset. In *INTERSPEECH*, pages 1521–1525.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE.

Christopher Gebauer, Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann. 2023. Exploiting diversity of automatic transcripts from distinct speech recognition techniques for children's speech. In *Proc. INTERSPEECH*, pages 4578–4582.

Igor Gitman, Vitaly Lavrukhin, Aleksandr Laptev, and Boris Ginsburg. 2023. Confidence-based Ensembles of End-to-End Speech Recognition Models. In *Proc. INTERSPEECH 2023*, pages 1414–1418.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang, editors. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Shahab Jalalvand, Matteo Negri, Daniele Falavigna, and Marco Turchi. 2015. Driving rover with segment-based asr quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1095–1105.

Aleksandr Laptev and Boris Ginsburg. 2023. Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 152–159. IEEE.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Florian Metze, Thomas Kemp, Thomas Schaaf, Tanja Schultz, and Hagen Soltau. 2000. Confidence measure based language identification. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1827–1830. IEEE.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. *Springer Handbook of Speech Processing*, pages 559–584.

David W Opitz and Richard F Maclin. 1997. An empirical evaluation of bagging and boosting for artificial neural networks. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 3, pages 1401–1405. IEEE.

Bohdan Pavlyshenko. 2018. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 255–258.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Holger Schwenk and Yoshua Bengio. 2000. Boosting neural networks. *Neural computation*, 12(8):1869–1887.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Fabio Valente. 2010. Multi-stream speech recognition based on dempster–shafer combination rule. *Speech Communication*, 52(3):213–222.

Pei-Hsin Wang, Sheng-Iou Hsieh, Shieh-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. 2020. Contextual temperature for language modeling.

Sinha Yamini and Siegert Ingo. 2021. Improving the accuracy for voice-assistant conversations in german by combining different online asr-api outputs. *Human Perspectives on Spoken Human-Machine Interaction*, pages 11–16.