

Speaker-Smoothed kNN Speaker Adaptation for End-to-End ASR

Shaojun Li, Daimeng Wei, Jiaxin GUO, ZongYao LI, Zhanglin Wu, Zhiqiang Rao, Yuanchang Luo, Xianghui He, Hao Yang

Huawei Translation Services Center

{lishaojun18, weidaimeng, shanghengchao, guojiaxin1, lizongyao, wuzhanglin2, raozhiqiang, luoyuanchang1, hexianghui, yanghao30}@huawei.com

Abstract

Despite recent improvements in End-to-End Automatic Speech Recognition (E2E ASR) systems, the performance can degrade due to vocal characteristic mismatches between training and testing data, particularly with limited target speaker adaptation data. We propose a novel speaker adaptation approach Speaker-Smoothed kNN that leverages k-Nearest Neighbors (kNN) retrieval techniques to improve model output by finding correctly pronounced tokens from its pre-built datastore during the decoding phase. Moreover, we utilize x-vector to dynamically adjust kNN interpolation parameters for data sparsity issue. This approach was validated using KeSpeech and MagicData corpora under in-domain and all-domain settings. Our method consistently performs comparably to fine-tuning without the associated performance degradation during speaker changes. Furthermore, in the all-domain setting, our method achieves state-of-the-art results, reducing the CER in both single speaker and multi-speaker test scenarios.

Index Terms: speech recognition, speaker adaptation, k-nearest neighbors

1. Introduction

Recently, end-to-end automatic speech recognition (E2E ASR) systems have demonstrated considerable performance improvements, aided by extensive training data amassed from a variety of speakers [1, 2]. Despite these advancements, the performance of E2E systems can plummet dramatically when confronted with significant voice character mismatch between training and testing conditions. In response to this issue, speaker adaptation algorithms seek to rectify the aforementioned mismatch by tailoring the ASR model to fit the specific characteristics of the target speaker.

The biggest challenge of speaker adaptation is that the adaptation data amount from the target speaker is usually very small. There are two types of methods to address such a challenge. The first type is model-based method, it studies how to effectively utilize few-shot speaker data. Some methods directly fine-tune the parameters of the pre-trained model [3], or train the model from speaker-independent parameters [4, 5]. To alleviate the overfitting problem caused by the limited adaptation data, the L2 norm [6], Kullback–Leibler divergence [7, 8] are introduced to regularize the adapted model. There are also many approaches utilizing data augmentation, for example, generating speaker data via TTS to augment the model [9]. However, most of these methods potentially suffer from significant performance drops when the amount of adaptation data decreases.

The second type is feature-based method, in which auxiliary speaker embeddings, such as i-vector [10] and x-vector

[11], are fed into an ASR model along with speech features. In this way, the model is working on the acoustic features, i.e. either by normalizing acoustic features to be speaker-independent [12, 13] or by introducing additional speaker related knowledge (e.g., i-vector) to adapt the acoustic model [14, 15]. A summary vector of each utterance can be trained to replace speaker i-vector [16]. To adapt to acoustic variability [17], shifting and scaling parameters are added in the layer-normalization layer. Recently, the attention mechanism is introduced to speaker-aware training (SAT). Typically, a speaker-aware persistent memory is incorporated to the transformer based ASR model [18, 19]. These memories performance always hit a bottleneck with the limited capacity of tens or hundreds of speaker embedding search space.

Considering the large size of E2E ASR training set, there may be utterances of similar voice characters as that of the target speaker. Thus, it is reasonable to utilize the similar utterances to be a supplement for target speaker data in the adaptation process. Motivated by this idea, we introduce a kNN retrieval techniques to ASR model which initially proposed for regulating autoregressive decoding (AD) of language model [20] and machine translation [21]. We expect this kNN classifier to find the correctly pronounced token from its pre-built datastore at each decoding to correct the model’s erroneous output. [22] modified kNN to apply to frame-level CTC decoding, but lacked further research on AD and the risk of false recall when data is sparse in kNN [23]. Therefore, we firstly explore the effectiveness of kNN token-level representation. It was found that there may be some problems when speaker adaptation is used directly on kNN, because kNN has the fixed interpolation parameters, T and λ . To solve this problem, we propose a Speaker-Smoothed kNN network, which uses extra information such as x-vector to dynamically adjust the interpolation ratio. [24] uses few-shot speaker data to extract similar pronunciations from training data, which is similar to our idea. However, this training method only uses utterance level information and re-training is required when the speaker is switched.

We test proposed Speaker-Smoothed kNN using KeSpeech and MagicData corpora with in-domain and all-domain settings. In the in-domain setting, our method performs close to fine-tuning without experiencing performance degradation like fine-tune and kNN method when the speaker changes. In the all-domain setting, our method achieves sota in both single speaker and speaker change scenarios, reduce 12.35 and 24.68 CER on the single speaker and multi-speaker test sets respectively.

2. Background

Here we give a brief introduction of kNN method. kNN uses token-level context retrieval to enhance the quality of a pre-

trained neural machine translation (NMT) model [25]. This method includes two steps:

Datastore Construction The datastore is a translation memory which converts bilingual sentence pairs into a set of key-value pairs. Given a reference corpus $(x, y) \in (\mathcal{X}, \mathcal{Y})$, the pre-trained NMT model generates the context representation $f_\theta(x, y_{<t})$ at each timestep t . Then we collect the output hidden state $f_\theta(x, y_{<t})$ as key and y_t as value to construct the whole datastore $(\mathcal{K}, \mathcal{V})$:

$$(\mathcal{K}, \mathcal{V}) = \bigcup_{(x, y) \in (\mathcal{X}, \mathcal{Y})} \{(f_\theta(x, y_{<t}), y_t), \forall y_t \in y\} \quad (1)$$

Inference with kNN Retrieval At the t -th decoding step, given the already generated words $\hat{y}_{<t}$, the current context representation $f_\theta(x, \hat{y}_{<t})$ is leveraged to generate a retrieval distribution $p_{kNN}(y_t | x, \hat{y}_{<t})$ over the entire vocabulary:

$$p_{kNN}(y_t | x, \hat{y}_{<t}) \propto \sum_{(h_i, v_i) \in N_t} \mathbb{I}_{y_t=v_i} \exp\left(\frac{-d(h_i, f_\theta(x, \hat{y}_{<t}))^2}{T}\right) \quad (2)$$

where the $d(\cdot, \cdot)$ stands for Euclidean distance function and T is the temperature to control the sharpness of softmax function. The final prediction distribution enhances vanilla NMT distribution p_{NMT} with the retrieval distribution p_{kNN} , and it is formally calculated as:

$$p(y_t | x, \hat{y}_{<t}) = \lambda p_{kNN}(y_t | x, \hat{y}_{<t}) + (1 - \lambda) p_{NMT}(y_t | x, \hat{y}_{<t}) \quad (3)$$

T and λ above are tuned interpolation coefficients.

3. Method

3.1. Preliminary Study

Given the limited research on the application of kNN to ASR and speaker adaptation, we initiated a preliminary study to demonstrate the potential of kNN in ASR. Utilizing Whisper, we obtained token-level representations denoted as h_i . We then sampled different speaker kNN representations across three dimensions, they are subdialect, speaker and token. The t-SNE results on h_i embeddings from Whisper are illustrated in Figure 1. A.x samples derive from the same subdialect, where A.1 samples the same tokens with varying colors representing different speakers, while A.2 includes multiple speakers, with different color schemes indicating different tokens. B.x settings, on the other hand, sample from different subdialects, choosing up to 200 representations for four tokens or speakers.

The A.x samples drawn from the same subdialect exhibit considerable overlap among speakers within the same token cluster (A.1), and a clear demarcation exists between different token clusters (A.2). These results align with our expectations and benefit our kNN retrieval. It is important to note the evident sparseness of the speaker adaptation data, as reflected in token cluster A.2 (token C), as the use of fixed temperatures and weights by kNN for interpolation increases the chance of error tokens.

B.x samples, collected from different subdialects, display overlaps between clusters in B.1 and boundaries between clusters in B.2, as anticipated. However, an unfortunate overlap is found in token C of B.2. We hypothesize that this overlap may be attributed to similar pronunciations of different tokens by speakers from diverse subdialects, in addition to data sparseness.

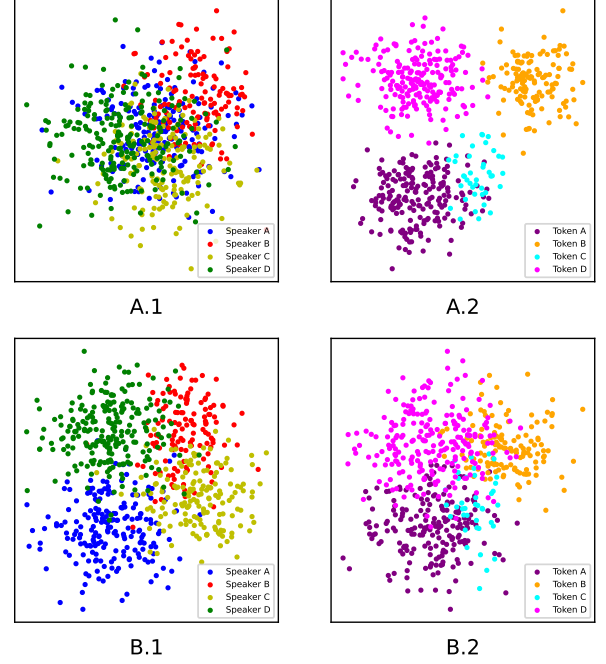


Figure 1: t-SNE result of the kNN datastore representation. A.x samples representation from the same subdialect, and B.x samples from different subdialects. Dots of different colors indicate different speaker or token clusters.

The above-mentioned overlaps, which kNN classification cannot distinguish effectively, are largely related to speaker accents. Therefore, the incorporation of speaker embedding into kNN classification may prove beneficial in distinguishing between these overlaps, thereby enhancing the accuracy of our approach.

3.2. Speaker-Smoothed kNN Network

Speaker recognition (SRE) models have been demonstrated to effectively distinguish between speakers [26, 27]. Speaker embeddings, derived from SRE models, have also been deployed directly for speaker adaptation [18, 19]. To address the challenges posed by sparsity and error overlap identified in our preliminary study, we now incorporate x-vector, a type of speaker information, into the kNN approach.

Figure 2 displays our process, starting with datastore construction as per Section 2. We use the SRE model to create an utterance-level x-vector, which we pair with its kNN value in a mapping relationship. Kster’s work [23] inspires us to design two estimators for variables T and λ dynamically. Unlike Kster, we simplify our network settings according to [28] and integrate x-vector. Noting the sparsity of our speaker data, we heed findings from [29], asserting networks utilizing distance and distinct token information yield better generalization results.

Concretely, at the t -th decoding step, we first retrieve K neighbors from the datastore, and consider the powers of 2 as the choices of k for simplicity. For each neighbor (h_i, v_i, e_i) , where e_i denoted the mapping x-vectors, we calculate the distance from the current context representation $d_i = d(h_i, f_\theta(x, \hat{y}_{<t}))$, as well as the count of distinct values in top i neighbors c_i . Denote $d = (d_1, \dots, d_K)$ as distances and $c = (c_1, \dots, c_K)$ as counts of values for all retrieved neigh-

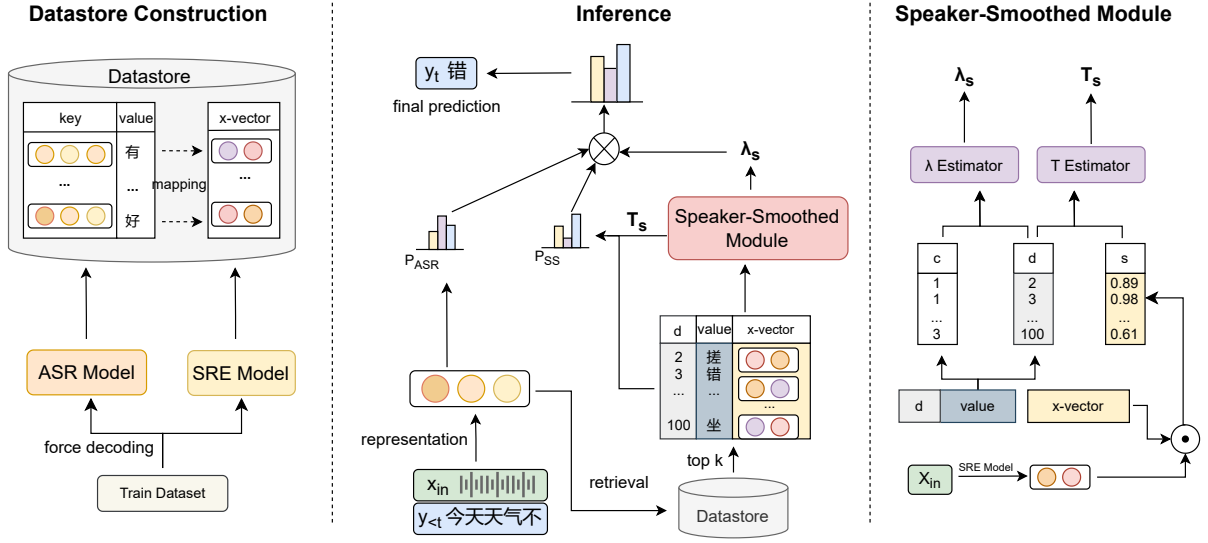


Figure 2: An overview of our proposed Speaker-Smoothed kNN framework

bors. Finally, we use input x to generate x -vector and calculate the dot product with e , here, denote $s = (s_1, \dots, s_K)$ as the x -vector similarity.

For dynamic distribution temperature T denoted as T_s , we concatenate d and s as input features to the T estimator network. From the preliminary study, the above operations are performed in the hope that the x -vector will help distinguish overlapping issues.

$$T_s = \exp(\mathbf{W}_1[d; s] + \mathbf{b}_1) \quad (4)$$

For dynamic λ denoted as λ_s , the mixing weight λ_s is computed by a d and c as inputs where $[\mathbf{W}_2; \mathbf{b}_2; \mathbf{W}_3; \mathbf{b}_3]$ are trainable parameters. These distance information can deal with the sparse problem from the preliminary study.

$$\lambda_s = \text{sigmoid}(\mathbf{W}_3 \text{ReLU}(\mathbf{W}_2[d; c] + \mathbf{b}_2) + \mathbf{b}_3) \quad (5)$$

It's noteworthy that our technique, in theory, can be employed with any autoregressive E2E ASR model, including RNN-T or Conformer. Once our network has been trained, it enables hot update in the speaker datastore during testing, thereby supporting on-the-fly adaptation.

3.3. Prediction and Training

In prediction phrase, we use our Speaker-Smoothed network to generate dynamic T_s and λ_s for each untranscribed token to utilize different T and λ . That is to say, we replace the fixed T in Equation 2 and λ in Equation 3 by λ_s and T_s varies at each predicting \hat{y}_t . Here p_{ASR} denotes ASR distribution and p_{SS} denotes Speaker-Smoothed kNN distribution

$$p(y_t | x, \hat{y}_{<t}) = \lambda_s p_{ASR}(y_t | x, \hat{y}_{<t}) + (1 - \lambda_s) p_{SS}(y_t | x, \hat{y}_{<t}) \quad (6)$$

For training, we fix the pre-trained ASR model and only optimize the Speaker-Smoothed kNN network by minimizing the cross entropy loss following Equation (6), which could be very efficient by only utilizing hundreds of training samples.

Subdialect	#Hours	#Million Tokens
All-Domain	860	11
In-Domain		
<i>Zhongyuan</i>	84	1.6
<i>Southwestern</i>	75	1.4
<i>Ji-Lu</i>	59	1.1
<i>Jiang-Huai</i>	46	0.9

Table 1: Number hours and tokens of in-domain and all-domain data collection in KeSpeech.

4. Experiment

4.1. Setup

We leverage the Whisper-medium pre-trained model for tracking speaker mismatches. This model demonstrates impressive efficacy on Mandarin Chinese tests. A pre-trained speaker recognition model aids [30] in extracting embeddings for similarity assessments and calculating variable s as per Section 3.2.

For training and kNN datastore construction, the KeSpeech¹ corpora is employed. We perform two categories of experiments based on datasets: (1) In-domain setting, we train exclusively with a dataset composed of a single subdialect; and (2) All-domain setting, utilizing the entire KeSpeech corpus - an open-source dataset of Mandarin and eight subdialects. We deploy phase-1 of KeSpeech, offering 895 hours of data from 34 cities across China. Table 1 details training data and datastore size.

Our evaluation employs two test sets based on speaker variation: (1) Single speaker test set. For this, we utilize two open-source datasets from the MagicData²: the Sichuan and Zhengzhou corpora, each representing Southwestern and Zhongyuan Mandarin dialects, respectively. These single speaker adaptation test sets consist of 20 speakers, each providing approximately 30 minutes of speech. The test sets are

¹<https://github.com/KeSpeech/KeSpeech>

²<https://magichub.com/datasets>

Method	Single Speaker		Multi Speaker	
	CER	Δ	CER	Δ
<i>Baseline</i>	29.17	NA	36.71	NA
In-Domain Setting				
<i>Fine-tune</i>	<u>17.29</u>	<u>11.88</u>	52.31	-15.60
<i>kNN</i>	17.92	11.25	47.5	-10.79
<i>Ours</i>	17.74	11.43	<u>19.18</u>	<u>17.53</u>
All-Domain Setting				
<i>Fine-tune</i>	18.92	10.25	13.05	23.66
<i>SAT</i>	18.2	10.97	14.26	22.45
<i>PAT</i>	17.66	11.51	12.81	23.90
<i>kNN</i>	18.11	11.06	16.47	20.24
<i>Ours</i>	17.54	11.63	12.3	24.41
<i>Fine-tune+Ours</i>	16.82	12.35	12.03	24.68

Table 2: Main Results in two settings. Δ means the difference from baseline.

allocated following [24], allocating 10 minutes for adaptation and 20 minutes for testing. (2) Multi-speaker test set. Here, we leverage the entire KeSpeech test set for evaluating performance in scenarios with varying speakers.

4.2. Implementation Details

For nearest neighbor retrieval, we construct a FAISS index [31]. We leverage inverted file system and product quantization for quick retrieval from large databases. Keys of examples are stored in fp16 format to conserve memory.

In training our Speaker-Smoothed network, the hidden size is set to 32. We directly use the KeSpeech dev set, training the network for about 4,000 steps. Our model is optimized with Adam, with a learning rate of $3e-4$ and a batch size of 32.

During inference, we set the beam search size to 5 and retain other parameters by default for all settings. For kNN retrieval, we set top k as 32, λ as 0.4 and T as 1000 for all experiments.

4.3. Main Results

Table 2 compares the proposed method with other adaptation methods in terms of CER on single and multi speakers. The Whisper baseline results showed that the CER performance was poor in both test sets, with mismatch.

In the in-domain experiment, fine-tuning yielded optimal results for single speaker adaptation. Yet, due to catastrophic forgetting, the multi-speaker test set experienced a CER increase. The kNN method’s significant CER decline on the single speaker test set denotes its effectiveness, but rise in the multi-speaker due to mismatching and inability to reject noise. Our Speaker-Smoothed kNN method, albeit slightly weaker than fine-tuning, excelled on the multi-speaker test set.

In the All-domain experiment, fine-tuning underperformed compared to its in-domain variant on the single speaker test set. Still, the multi-speaker test set’s CER significantly reduced compared to the baseline, owing to a complete match. We implemented speaker-aware training (SAT) [15], applicable to various ASR models, and PAT [24], learning speaker adaptation from the training dataset, for comparison, maintaining the same supervised settings. Both SAT and kNN fell short of fine-tuning on the multi-speaker test set. Our proposed method, however,

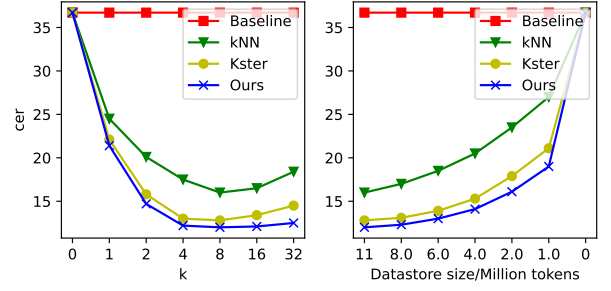


Figure 3: Left side is the CER trend of different methods when use different top k , right side is the CER trend in different sample datasore size. Both experiment done on the multi-speaker test of all-domain setting.

outperformed on both single and multi-speaker test sets, enhancing further when used continuously on a fine-tuning base.

4.4. Analysis

In the all-domain experimental setting with a multi-speaker test set, as shown on the left of Figure 3, the top k types of noise increase gradually as k expands, particularly for sparse data. Compared to the direct application of kNN, and Kster [23], which employ a similar smoothed network without an x-vector, performance diminishes to various extents. As displayed on the right of Figure 3, as the datasore contracts, the noise volume incrementally escalates, and our method exhibits a more gradual decline. These two experiments underscore that our Speaker-Smoothed method handles error recalls more effectively.

4.5. Inference Cost

The inference expense of kNN memorization retrieval, discussed in previous studies [28, 32], may result in reduced speed as the datasore expands. This study explores this on the all-domain setting based on Whisper-medium. With a batch size of 16, we find that the average inference speed is 87.1% of that of the kNN-free method (given that the weight of the SRE model is relatively light, its consumption was disregarded). In practice, this speed reduction is acceptable as it represents a balancing act between performance and processing time. To achieve superior inference speed, we could potentially replace it with other memorization-retrieval variants [28, 33]. This is a prospect for future work.

5. Conclusion

Our research presents a novel approach to enhancing E2E speaker adaptation performance through Speaker-Smoothed kNN, notably in situations of limited adaptation data. By utilizing x-vector information, we achieve dynamic adjustment of interpolation ratios, leveraging the similarity in voice characters of training data. We attain considerable improvement, outperforming established techniques like fine-tuning and speaker-aware training in the KeSpeech and MagicData test sets. By demonstrating the compelling application of kNN in ASR speaker adaptation, we pave the way for future inquiries into this promising area.

6. References

- [1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, p. 1240–1253, Dec 2017. [Online]. Available: <http://dx.doi.org/10.1109/jstsp.2017.2763455>
- [2] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S. yin 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. Shang-guan, Y. Sheth, T. Strohmaier, 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," 2020.
- [3] Y. Huang, G. Ye, J. Li, and Y. Gong, "Rapid speaker adaptation for conformer transducer: Attention and bias are all you need," in *Interspeech 2021*, Aug 2021. [Online]. Available: <http://dx.doi.org/10.21437/interspeech.2021-1884>
- [4] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 171–176.
- [5] X. Xie, X. Liu, T. Lee, and L. Wang, "Bayesian learning for deep neural network adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2096–2110, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:229156237>
- [6] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013. [Online]. Available: <http://dx.doi.org/10.1109/icassp.2013.6639212>
- [7] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KI-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013. [Online]. Available: <http://dx.doi.org/10.1109/icassp.2013.6639201>
- [8] Z. Meng, Y. Gaur, J. Li, and Y. Gong, "Speaker adaptation for attention-based end-to-end speech recognition," in *Interspeech*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202719968>
- [9] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. S. Koppula, and O. Tuzel, "Text is all you need: Personalizing asr models using controllable speech synthesis," 2023.
- [10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018. [Online]. Available: <http://dx.doi.org/10.1109/icassp.2018.8461375>
- [12] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2011.
- [13] N. Tomashenko and Y. Estève, "Evaluation of feature-space speaker adaptation for end-to-end acoustic models," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. [Online]. Available: <https://www.aclweb.org/anthology/L18-1500>
- [14] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.
- [15] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [16] K. Vesely, S. Watanabe, K. Zmolikova, M. Karafiat, L. Burget, and J. H. Cernocky, "Sequence summarizing neural network for speaker adaptation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5315–5319.
- [17] T. Kim, I. Song, and Y. Bengio, "Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, 2017. [Online]. Available: <https://arxiv.org/abs/1707.06065>
- [18] L. Sari, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised speaker adaptation using attention-based speaker memory for end-to-end asr," 2020.
- [19] Y. Zhao, C. Ni, C.-C. Leung, S. Joty, E. S. Chng, and B. Ma, "Speech Transformer with Speaker Aware Persistent Memory," in *Proc. Interspeech 2020*, 2020, pp. 1261–1265.
- [20] U. Khandelwal, O. Levy, and D. J. et al., "Generalization through memorization: Nearest neighbor language models," in *ICLR*, 2020.
- [21] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Nearest neighbor machine translation," in *ICLR*, 2021.
- [22] J. Zhou, S. Zhao, Y. Liu, W. Zeng, Y. Chen, and Y. Qin, "Knn-ctc: Enhancing asr via retrieval of ctc pseudo labels," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 006–11 010.
- [23] Q. Jiang, M. Wang, J. Cao, S. Cheng, S. Huang, and L. Li, "Learning kernel-smoothed machine translation with retrieved examples," 2021.
- [24] Y. Gu, Z. Du, S. Zhang, Q. Chen, and J. Han, "Personality-aware Training based Speaker Adaptation for End-to-end Speech Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1249–1253.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.
- [26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [27] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," 2021.
- [28] Y. Dai, Z. Zhang, Q. Liu, Q. Cui, W. Li, Y. Du, and T. Xu, "Simple and scalable nearest neighbor machine translation," in *ICLR*, 2023.
- [29] X. Zheng, Z. Zhang, J. Guo, S. Huang, B. Chen, W. Luo, and J. Chen, "Adaptive nearest neighbor machine translation," in *ACL*, 2021.
- [30] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *arXiv: Audio and Speech Processing*, arXiv: Audio and Speech Processing, Dec 2020.
- [31] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021. [Online]. Available: <https://doi.org/10.1109/TBDATA.2019.2921572>
- [32] Y. Du, W. Wang, Z. Zhang, B. Chen, T. Xu, J. Xie, and E. Chen, "Non-parametric domain adaptation for end-to-end speech translation," in *EMNLP*, 2022.
- [33] Y. Meng, X. Li, X. Zheng, F. Wu, X. Sun, T. Zhang, and J. Li, "Fast nearest neighbor machine translation," in *Findings of ACL*, 2022.