Science Research
2025, Vol. 13, No. 2, pp. 13-20
https://doi.org/10.11648/j.sr.20251302.11

Research Article

Korean Spoken Accent Identification Using T-vector Embeddings

Yong Su Om*, Hak Sung Kim

Institute of AI Technology, University of Science, Pyongyang, Democratic People's Republic of Korea

Abstract

In this paper, we introduce a spoken accent identification system for the Korean language, which utilize t-vector embeddings extracted from state-of-the-art TitaNet neural network. To implement the Korean spoken accent identification system, we propose two approaches: First, we introduce a collection method of training data for the Korean spoken accent identification. Korean accents can be broadly classified into four categories: standard accent, southern accent, northwestern accent and northeastern accent. Generally, in Korean language, the speech data for standard accent can be easily obtained via different videos and websites, but the rest of the data except standard accent are very rare and therefore difficult to collect. To mitigate the impact of this data scarcity, we introduce a synthetic audio augmentation using Text-to-Speech (TTS) synthesis techniques. This process is done under the condition that the synthetic audio generated by TTS should be retain accent information of original speaker. Second, we propose an approach to build the deep neural network (DNN) for Korean spoken accent identification in a manner that fine-tune the trainable parameters of a pre-trained TitaNet speaker recognition model by using aforementioned training dataset. Based on the trained TitaNet model, the accent identification is performed using t-vector embedding features extracted from that model, and cosine distance function. The experimental results show that our proposed accent identification system is superior to the systems based on other state-of-the-art DNNs such as the x-vector and ECAPA-TDNN.

Keywords

Spoken Accent Identification, T-vector, Deep Neural Network, Fine-tuning, Synthetic Audio Augmentation (SSA), Titanet, 1D time -channel Separable Convolution, Squeeze and-Excitation (SE)

1. Introduction

Accent refers to different ways of pronouncing a language within a community [1]. Spoken accent identification is a family of spoken language identification, which is used as an essential technology in various areas of social life, including automatic speech recognition, video classification, and customer services based on user-agent voice commands.

Currently, the main research in the field of spoken language recognition has been focused on distinguishing closely related languages, dialects and accents correctly, provided that the recognition techniques are mature enough to distinguish between essentially distinct languages (e.g., Korean, Chinese, and English). Caused by more subtle acoustic and linguistic variations, accent and dialect identification are generally more difficult than language identification [8].

*Corresponding author: oys1978@star-co.net.kp (Yong Su Om)

Received: 18 May 2025; **Accepted:** 12 June 2025; **Published:** 30 June 2025

Early, statistical techniques such as GMM-UBM [2], Joint Factor Analysis (JFA) [3], and i-vectors [1, 4, 8] were widely employed in language and accent identification. In particular, S. Ramoji et al. [32] proposed a fully supervised learning-based i-vector modeling method (s-vector) where class labels of speech recordings are directly introduced into the i-vector model, and through experiments, they confirmed that it has superior performance compared to the traditional i-vector method.

After that, with the rapid development of deep learning techniques, spoken language identification has been converted from statistical techniques to DNN based methods such as x-vectors [5, 21] and ResNet [6], and its performance has improved significantly. Language and accent identificaition using deep learning can be broadly divided into end-to-end neural network based methods and DNN-embedding feature based methods, the latter approach being mainstream. Deep embedding features are fixed-dimensional representations of speech utterances extracted from DNNs, among which x-vector and their variants (e.g., F-TDNN [10], E-TDNN [11]) have been widely used. An ideal DNN-embedding maximizes inter-class variations and minimizes intra-class variations [9].

The widespread use of x-vector systems has led several approaches for extracting more discriminative deep-embedding features, such as attention-based methods [12, 15], residual connection based methods [13], Neural i-vectors based on squeeze-excitation modules [14], and transformer-based methods [7]. However, these methods still suffer from the performance degradation due to external environmental conditions such as background noise, channel distortion, and room reverberation.

Recently, ECAPA-TDNN model proposed by B. Desplanques et al. [16] has achieved outstanding performance in language and speaker recognition tasks as an advanced neural network architecture that eliminated some limitations of TDNN based x-vector model [17-19].

This model emphasized the functionality of frame-level block and statistics pooling layer using modeling techniques, such as 1-Dimensional SE-Res2Block to explicitly model channel interdependence from input acoustic feature data, multi-layer feature aggregation to combine different hierarchical frame-level features, and statistics pooling with channel-dependent frame attention. As a result, the quality of the embedding features extracted from this model is improved.

Meanwhile, N. R. Koluguri et al. [20] proposed a novel neural network architecture, Titanet (also called t-vector), using one-dimensional depth-wise separable convolution, squeeze-excitation (SE) layers with global context, and statistics pooling layer based on channel attention. The authors demonstrated that this deep neural model with scalable neural network architecture achieves remarkable performance in speaker recognition and speaker diarization while reducing the parameter count compared to ECAPA-TDNN model.

However, advanced deep models such as ECAPA-TDNN and Titanet can deliver state-of-the-art performance only when sufficient amount of training data is available. Moreover, as the neural network architecture becomes complex and the parameter count increases, more training data is required to avoid overfitting in the model's training and to ensure the stability of the model. In particular, the training set should be consisted of a balanced amount of data from each class.

To solve this problem, researchers have developed several approaches including SpecAugment, additive noise augmentation from the MUSAN dataset, reverberation noise augmentation from the RIRs dataset, speed perturbation, voice conversion, and the combination of various signal processing techniques.

Recently, in order to extract more discriminative neural embeddings, researchers have proposed various methodologies, a widely used strategy [22, 25] involves leveraging acoustic and linguistic features extracted from sub-blocks of large pre-trained models trained on extensive corpora (e.g. Whisper [23], XLS-R [33], Wav2vec2 [24], and WavLM [31]). The incorporation of these pre-trained models provides a strong basis for acoustic feature extraction and DNN-embedding learning, thus mitigating the performance deterioration caused by lack of training data in the low-resource language domain. However, the drawback of this approach is that the inference speed is slow due to increased computational complexity of the neural network.

In this paper, we propose a method for Korean spoken accent identification using t-vector embeddings extracted from state-of-the-art Titanet model. The accent of Korean speech differs slightly according to the regional characteristics, which can be divided into four major categories: standard accent, southern accent, northwestern accent and northeastern accent.

First of all, we propose a text-speech synthesizer-based data augmentation approach to prevent performance degradation of the learning model caused by lack of training data in the Korean spoken accent identification task.

Next, we propose a method to build the learning model for Korean spoken accent identification by fine-tuning the parameters of a pre-trained Titanet speaker recognition model using the aforementioned training data set, and demonstrate its effectiveness from experimental results.

The rest of this paper is organized as follows. Section 2 describes the data collection for Korean spoken accent identification and the synthetic speech augmentation using GradTTS. We also describe the architecture of TitaNet neural network, model learning and the classifier applied for Korean spoken accent identification. The experimental results and analysis are presented in Section 3 and finally Section 4 concludes.

2. Research Method

2.1. Dataset

In this section a speech corpus for Korean spoken accent

identification is introduced. We performed the audio data collection for four accents of Korean language: Standard accent, Southern accent, Northwest accent and Northeast accent. The standard accent data in Korean speech can be easily obtained because most of the video such as news or film is composed of standard accent speech. However, the rest of the data except standard accent are very rare and therefore difficult to collect.

The audio data was collected from various kinds of videos (films, interviews, lectures, news, etc.) downloaded from Korean Central TV channels and different websites. We extracted audio signals into 1-channel wav format with 16 KHz sampling rate from each video using FFMPEG module. Energy-based Voice Activation Detection (VAD) was applied for removal of silence frames in the training data.

Each speech segment contains only one speaker and different environments such as indoor, street, telephone, etc. Then, each of the resulting speech segments was manually reviewed to remove files containing nuisance information (e.g., music, other languages, laughter). The general specification of the collected training dataset are listed in Table 1.

Table 1. *General specification of the collected training data set.*

Accent	Segment Number	Speaker Number	Min length (s)	Max length (s)	Avg length (s)	Total (hrs)
Normal	17372	104	2.5	18.5	9.6	48.2
Southern	2720	17	2.6	14.3	8.6	6.5
Northwest	4018	28	2.2	12.6	10.2	11.4
Northeast	1838	13	3.4	12.8	9.4	4.8
Total	25948	162	2.7	14.5	9.5	70.9

2.2. Data Augmentation Using Text-to-speech Synthesis

Preparing sufficient speech resources for each accent is crucial for DNN based accent identification task. Generally, in Korean language, standard accent utterance are readily available, but utterance of other three accents are relatively scarce.

We employ synthetic audio augmentation (SAA) using text-to-speech synthesis techniques to mitigate the lack of training data. In this work, we implement SAA leveraging a Korean text-to-speech synthesis system that was already developed based on Grad-TTS [26]. Grad-TTS is an acoustic feature generator based on denoising diffusion probabilistic model that has high potential in modeling complex data distributions.

The basic principle of diffusion probabilistic model is to build a forward diffusion process by iteratively destroying original data until some simple distribution (usually standard normal distribution) is obtained, and then to build a backward diffusion by learning the neural network so that it follows the trajectories of reverse time forward diffusion.

The acoustic feature generator is built by training the model so that noise is sequentially transformed into mel-spectrogram extracted from the training sample. This model provides explicit control of inference speed and is capable of generating mel-spectrograms of better quality compared to tacotron2. We use HiFi-GAN vocoder [27] to generate synthetic audio from Mel-spectrogram.

We obtained 16-hour training audio data from one speaker with clean voice with respect to each accent (except for standard accent) in a noise-free indoor environment. Here, the text data collected for model learning consisted of sentences containing more than 3000 most frequent words, each of which consisted of 10 to 15 words. Then, we built an acoustic feature generator and a HiFi-GAN vocoder model for speech synthesis using the dataset composed of text-speech pairs. By leveraging the trained speech synthesis models, we obtained the synthetic audio of 24 hours for each accent from pre-prepared plain-text data.

We have confirmed that synthetic audio generated by this method retains almost the accent information of the original speaker as MOS 4.32. Consequently, the amount of training data for three accents, namely, Southern accent, Northwest accent and Northeast accent, is 42.5, 47.4 and 40.8 h, including the data used for speech synthesis, 36 h (training data 12 h, synthetic speech 24 h), respectively.

The resulting total dataset is separated by a ratio of 8:1:1, of which 80% are used for tuning the hyperparameters of pre-trained TitaNet model, 10% for training the t-vector classifier, and 10% for classification testing.

2.3. TitaNet Model

In this study, we implement the Korean spoken accent identification system using TitaNet model as a baseline. TitaNet is a neural network adopted encoder architecture of the

ContextNet ASR model [28], which yields better performance over other state-of-the-art models in speaker recognition by leveraging 1D time depth-wise channel separable convolution, Squeeze and-Excitation (SE) layers with global context, and channel attention based statistics pooling layer.

As depicted in Figure 1, the model consists of a ContextNet-B×R×C model encoder and attention pooling decoder, where B is the number of blocks, R is the number of repeated sub-blocks per block, and C is the number of filters in the convolution layers of each block.

Prologue block B_0 and epilogue block B_N consist of the same components: one-dimensional convolution, batch normalization, and ReLU layers. The kernel size of filter in prologue and epilogue blocks are 3 and 1, respectively. These blocks do not contain residual connections and dropout layers. Each intermediate block begins with time-channel separable convolution layer (stride size: 1, dilation size: 1) followed by batch normalization, relu, and dropout layers. Each time-channel separable module consists of two parts: a depth-wise convolutional layer and a pointwise convolutional layer.

Depth-wise convolutions apply a single filter per input channel, and Pointwise convolutions are 1×1 convolutions, which create a linear combination of the outputs of the depth-wise layer.

These layers are repeated R times, which can be modified to vary the depth of the network. These repeated layers are residually connected with Squeeze and Excitation layers with global average pooling for context inclusion. The width of the network can be scalable by varying output channel filter sizes of each middle block. The width and depth of TitaNet model are easily changed by varying these filter sizes, C and the number of repeated layers, R respectively.

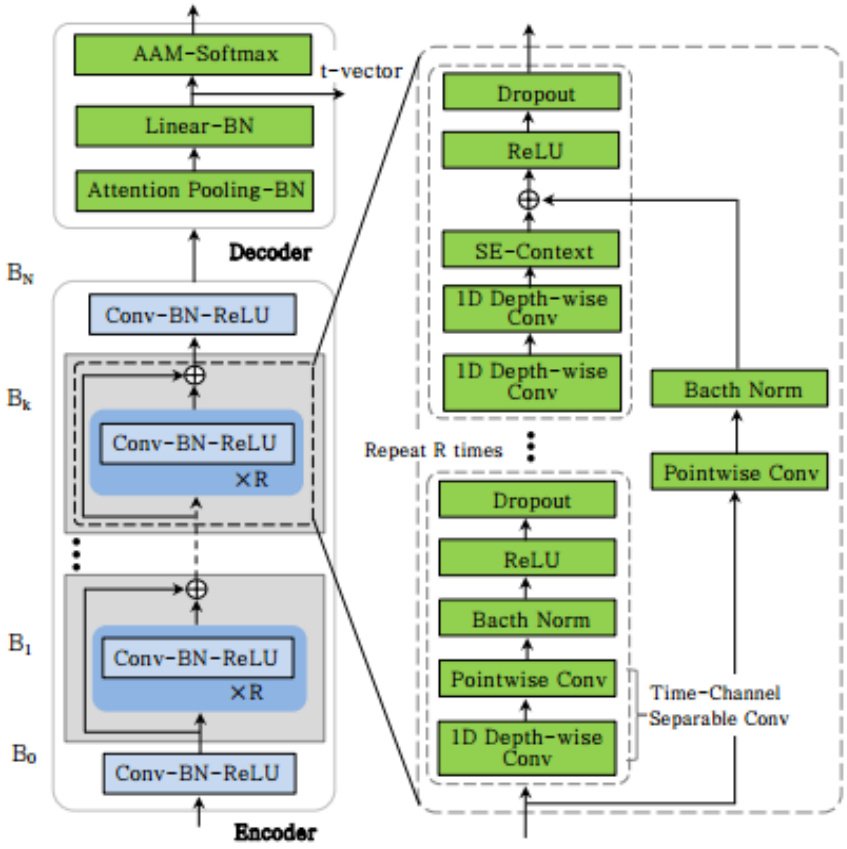

Figure 1. *TitaNet model Architecture.*

The top level acoustic features obtained from encoder's epilogue block is fed into an attentive statistics pooling layer, where a time-independent intermediate feature S of size B (minibatch size) × 3072 is obtained. This intermediate feature S are passed through two linear layers whose output sizes are 192 and N (number of classes to classify), respectively, and finally, the probability of each class is computed from the training data. The fixed-length embedding feature (t-vector) is extracted as 192 dimensional vector before the final logits linear layer.

2.4. Model Learning

In this study, as a baseline for implementing Korean spoken accent identification system, we use a pre-trained TitaNet-S speaker recognition model, which was trained by us-

ing Voxceleb1 dataset. The VoxCeleb1 dataset used for the model learning consist of human speech segments extracted from interview videos uploaded to YouTube, which contains more than 150,000 utterance segments of 240 hours, including more than 1200 speakers. Based on this baseline model, we generate a deep learning model that can identify four accent of Korean language by fine-tuning the pre-trained model using the training dataset mentioned in Section 2.1.

Fine-tuning is taking a pre-trained model and further training it on a smaller dataset specific to a particular task or application. Applying of this training strategy aims to avoid the overfitting phenomenon caused by data lack in Titanet model training which requires a large number of training data, and to improve the accuracy of the model.

The input acoustic features are 80 dimensional Mel-spectrograms computed using a 512 FFT and a Hann window. We compute the acoustic features with frame length 25 ms and shift 10 ms in the frequency range of 20-7800 Hz. Next, Mel-spectrogram features are normalized with respect to the frequency axis.

The baseline model was trained for 16 epochs with mini-batch size of 64 on a GeForce RTX 2070 GPU computer. During model training, we employs SGD optimizer and initial learning rate (LR) 0.01 using cosine annealing LR scheduler. Subsequently, the baseline model is fine-tuned for 9 epochs using the dataset for Korean spoken accent identification. In AAM-Softmax loss function applied to minimize intra-class variance, the margin m is set to 30 and the scale s is set to 0.2. The NVIDIA NeMo toolkit Open Source [30] was used for model building.

2.5. Classifier

We extract t-vector embedding features from the classifier's train data using the trained model. The resulting t-vectors are centered using the mean of training data. Subsequently, we applied linear discriminant analysis (LDA) process not reducing the dimensionality of the data.

For the t-vector W_{test} of the test accent inferred from the trained model and the t-vectors W^{a}_{target} of the target accent, the cosine score is computed as follows:

$$score\left(W_{test}, W^{a}_{target}\right) = \frac{\bar{W}^{T}_{test} \cdot \bar{W}^{a}_{target}}{\left\|\bar{W}_{test}\right\|\left\|\bar{W}^{a}_{target}\right\|} \tag{1}$$

Where,

$$\bar{W}_{test} = A^{T} W_{test} \tag{2}$$

and A is the LDA transformation matrix.

Further, The $\bar{W}^{a}_{target}$ is the mean of t-vectors extracted from all the training utterances in accent a, which is calculated as follows:

$$\bar{W}^{a}_{target} = \frac{1}{N_a}\sum_{i=1}^{N_a}\bar{W}^{a}_{i} \tag{3}$$

where N_a is the number of training utterances in accent a and $\bar{W}^{a}_{i}$ is the t-vector of the i-th enrollment utterance of accent a projected into the LDA transform space by Eq. (2).

For performance comparison with another classifier, in addition to cosine distance, we also use Gaussian classifier.

Given the t-vector W_{test} of a test utterance, the log-likelihood of a target accent a is computed as follows:

$$\Lambda_{W_{test}} = \bar{W}^{T}_{test}\Sigma^{-1}m_a - \frac{1}{2}m_a^{T}\Sigma^{-1}m_a \tag{4}$$

where m_a is the mean vector of accent a and Σ is the covariance matrix shared across all accents. It is computed as follows:

$$\Sigma = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_a}\sum_{i=1}^{N_a}\left(\bar{W}^{a}_{i} - m_a\right)\left(\bar{W}^{a}_{i} - m_a\right)^{T} \tag{5}$$

where

$$m_a = \frac{1}{N_a}\sum_{i=1}^{N_a}\bar{W}^{a}_{i} \tag{6}$$

and K is the number of total accents.

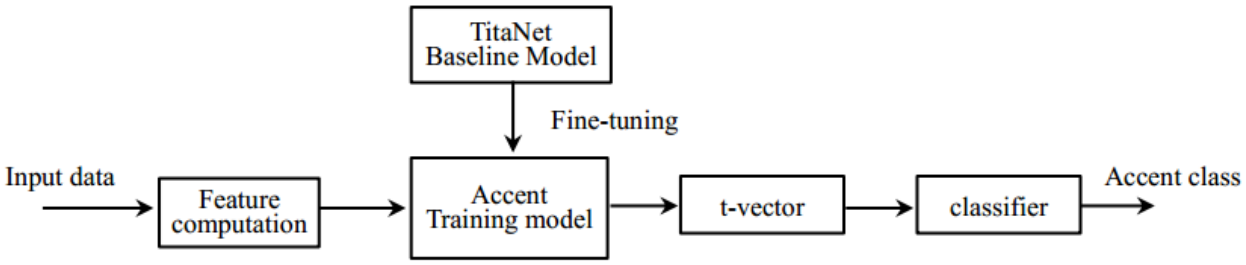

Figure 2. *The configuration of the Korean spoken accent identification system using t-vector embeddings.*

3. Results and Anaysis

In this section, we evaluate the performance of the accent identification system based on the method described in Section 2 and analyze the experimental results. The system configuration is shown in Figure 2.

In the experiments, the performance of Korean spoken accent identification system is reported in terms of equal error rate (EER). EER indicates the operating point on detection error trade-off (DET) curve at which false alarm and miss rates are equal. EER per target accent is computed in a manner that other accents serve as non-target trials.

We compared the discrimination performance between TitaNet and two typical models after training them in two stages of pre-training and fine-tuning as described in Section 2. For fair comparison, all of the neural networks were trained by AAM-Softmax loss function and their performance was evaluated by Cosine classifier. Table 2 shows the experimental results.

Table 2. *General specification of the collected training data set.*

Model	Number of model parameters (M)	EERavg (%)
x-Vector [5]	4.6	4.54
ECAPA-TDNN [16]	6.2	1.16
TitaNet-S [20]	6.4	0.93

As can be noticed in Table 2, the average EER of TitaNet-S model is the lowest. This can be attributed to the superior characteristics of 1D time-channel separable convolution and squeeze-excitation module involved in the architecture of TitaNet model. The networks based on 1D time-channel separable convolutions is fast, trains well, therefore that is widely used in practice [29].

Next, we evaluated the discrimination performance for each individual accent in the Korean spoken accent identification system using the Titanet-S model.

Table 3. *Equal error rate (EER) for each individual accent in the proposed Korean spoken accent identification system.*

Accent	EER (%)
Standard	0.43
Southern	0.75
Northwest	1.62
Northeast	0.94

As shown in the table above, for standard accents, data augmentation using speech synthesis was not conducted, but the EER was lowest, with 0.43%. This can be attributed to the diversity of training data for standard accent that contain relatively a large numbers of speakers.

Also, the EER is 0.75% in Southern accent, 1.62% in Northwest accent and 0.94% in Northeastern accent. In particular, in training dataset for each individual accent, synthetic audio accounts for more than 50% of that total amount. This means that the synthetic audio generated by TTS retain almost the original speaker's accent information and thus positively affect on the performance of Korean spoken accent identification system.

Next, we evaluated the performance for each target accent in terms of two distinct classifiers.

Table 4. *The Performance for each target accent in terms of two classifiers.*

Accent	Classifier Scoring	
	Cosine Scoring	Gaussian Scoring
Standard	0.43	0.45
Southern	0.75	0.83
Northwest	1.62	1.69
Northeast	0.94	1.12

From the above results, it can be seen that cosine classifier provides better performance than Gaussian classifier in terms of all of the accents.

This can be considered as the t-vector is extracted from the TitaNet-S model trained by AAM-Softmax loss function, which minimizes the intra-class variance in the angular space.

Finally, we performed a more intuitive analysis using confusion matrix based on top-1 classification accuracy.

Table 5. *Confusion Matrix based on top-1 classification accuracy.*

		Actual			
		Standard	Southern	Northwest	Northeast
Predicted	Standard	99.2	0.1	0.5	0.1
	Southern	0.4	98.8	0.7	0.2
	Northwest	1.3	0.7	97.8	0.1
	Northeast	0.8	0.3	0.5	98.3

As can be noticed from the diagonal values in the table 5,

in all cases, most of the results correspond to the correct accent class. The greatest confusion with respect to standard accent was found with 0.5% in the northwest accent. Specifically, the greatest confusion in testing dataset was found with 1.3% between Northwest and standard accents. This may be due to the fact that the two accents are geographically adjacent in terms of the area of use. The classification accuracy for standard accent is the highest, with 99.2%. This can explain by the fact that the diversity of speakers is relatively high in the training data for standard accent.

4. Conclusion

In this paper, we proposed a Korean spoken accent system by using t-vector embedding features and demonstrated its effectiveness via experiments. First, we introduced a training corpus for Korean spoken accent identification. In particular, we introduced a synthetic audio augmentation for mitigating the lack of training data, and confirmed that synthetic speech by TTS preserves the original speaker's accent information and consequently has a good effect on the performance of the accent classification model. Next, we built a Korean accent identification model by fine-tuning the pre-trained TitaNet speaker model and achieved EER of 0.43%, 0.75%, 1.62%, and 0.94% for 4 Korean accents, i.e., standard, southern, northwestern, and northeast accent, respectively.

This is the first report on the implementation of the Korean spoken accent identification. In future search, we will investigate in more depth the effects of other factors, like gender, age etc. on the accent identification task.

Abbreviations

DNN	Deep Neural Network
SAA	Synthetic Audio Augmentation
TTS	Text To Speech
VAD	Voice Activation Detection
ASR	Automatic Speech Recognition
SE	Squeeze and-Excitation
LR	Learning Rate
EER	Equal Error Rate

Author Contributions

Yong Su Om: Investigation, Writing-review & editing, Software.

Hak Sung Kim: Methodology, Validation.

All authors read and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] H. Behravan, V. Hautamaki et al., "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish," Speech Communication, vol. 19, pp. 118-129, 2015. http://dx.doi.org/10.1016/j.specom.2014.10.004

[2] Sreedhar Potla, Vishnu Vardhan. B, "Spoken Language Identification using Gaussian Mixture Model-Universal Background Model in Indian Context," International Journal of Applied Engineering Research, vol. 13, no. 5, pp. 2694-2700, 2018. https://www.ripublication.com/ijaer18/ijaerv13n5_79.pdf

[3] 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. http://dx.doi.org/10.1109/ICComm.2018.8453731

[4] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, R. Dehak, "Language Recognition via Ivectors and Dimensionality Reduction," INTERSPEECH, 2011, pp. 857–860. http://dx.doi.org/10.21437/Interspeech.2011-328

[5] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken Language Recognition using Xvectors," pages 105–111, June 2018a. http://dx.doi.org/10.21437/Odyssey.2018-15

[6] He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, in 'Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)', pp. 770–778. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_ResidualLearning_CVPR_2016_paper.pdf

[7] Katta, S. V., Umesh, S. et al. (2020), 'S-vectors: Speaker embeddings based on transformer's encoder for text-independent speaker verification', arXiv preprint arXiv: 2008.04659 http://dx.doi.org/10.48550/arXiv.2008.04659

[8] Behravan, H., Hautama ̈ki, V., Kinnunen, T., 2013. Foreign accent detection from spoken Finnish using i-Vectors. In: INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25–29, pp. 79–83. http://dx.doi.org/10.21437/Interspeech.2013-42

[9] S. Soonshin et al., "Self-Attentive Multi-Layer Aggregation with Feature Recalibration and Deep Length Normalization for Text-Independent Speaker Verification System", Electronics 2020, 9, 1706; https://doi.org/10.3390/electronics9101706

[10] M. Rouvier et al, "Review of different robust x-vector extractors for speaker verification," EUSIPCO 2020, pp. 366-370 http://dx.doi.org/10.23919/Eusipco47968.2020.9287426

[11] D. Povey, G. Cheng, Y. Wang, et al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH 2018, Hyderabad, India, sep 2018. http://dx.doi.org/10.21437/Interspeech.2018-1417

[12] Wang, Q., Okabe, K., Lee, K. A., Yamamoto, H. & Koshinaka, T. (2018), Attention mechanism in speaker recognition: What does it learn in deep speaker embedding?, in '2018 IEEE Spoken Language Technology Workshop (SLT)', IEEE, pp. 1052–1059. https://doi.org/10.3390/app13116410

[13] Yanpei Shi, "Improving the Robustness of Speaker Recognition in Noise and Multi-Speaker Conditions Using Deep Neural Networks," PhD Thesis, Department of Computer Science, University of Sheffield, 2021. https://etheses.whiterose.ac.uk/29662/

[14] Ville Vestman, Kong Aik Lee, Tomi H. Kinnunen, "Neural i-vectors", Odyssey 2020 The Speaker and Language Recognition Workshop, 2020, pp 67–74. http://dx.doi.org/10.21437/Odyssey.2020-10

[15] India Massana, M. A., Safari, P. & Hernando Peric´as, F. J. (2019), Self multi-head attention for speaker recognition, in 'Interspeech 2019: the 20th Annual Conference of the International Speech Communication Association: 15-19 September 2019: Graz, Austria', International Speech Communication Association (ISCA), pp. 4305–4309. https://doi.org/10.48550/arXiv.1906.09890

[16] B. Desplanques, et al., "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," arXiv preprint arXiv: 2005.07143, 2020. https://doi.org/10.48550/arXiv.2005.07143

[17] Hengyi Zou, Sayaka Shiota, "Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation Considering Speaker Variability for Speaker Verification," Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024.

[18] Vishwas M. Shetty et al., "ENHANCING AGE-RELATED ROBUSTNESS IN CHILDREN SPEAKER VERIFICATION," arXiv: 2502.10511v1 [eess. AS] 14 Feb 2025. https://doi.org/10.48550/arXiv.2502.10511

[19] Li Zhang et al., "Adaptive Data Augmentation with NaturalSpeech3 for Far-field Speaker Verification," arXiv: 2501.08691v1 [cs. SD] 15 Jan 2025 https://doi.org/10.48550/arXiv.2501.08691

[20] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 8102–8106. https://doi.org/10.48550/arXiv.2110.04410

[21] Ahmad Iqbal Abdurrahman, Amalia Zahra, "Spoken language identification using i-vectors, x-vectors, PLDA and logistic regression", Bulletin of Electrical Engineering and Informatics Vol. 10, No. 4, August 2021, pp. 2237~2244 http://dx.doi.org/10.11591/eei.v10i4.2893

[22] S. K. Gupta, S. Hiray, P. Kukde, "Spoken Language Identification System for English-Mandarin Code-Switching Child-Directed Speech", INTERSPEECH 2023, pp. 4114~4118 https://doi.org/10.48550/arXiv.2306.00736

[23] Alec Radford et al., "Robust speech recognition via largescale weak supervision," in International Conference on Machine Learning. PMLR, 2023, pp. 28492–28518. https://doi.org/10.48550/arXiv.2212.04356

[24] Alexei Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," Ad vances in neural information processing systems, vol. 33, pp. 12449–12460, 2020. https://doi.org/10.48550/arXiv.2006.11477

[25] Amrutha Prasad et al., FINE-TUNING SELF-SUPERVISED MODELS FOR LANGUAGE IDENTIFICATION USING ORTHONORMAL CONSTRAINT," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11921-11925. http://dx.doi.org/10.1109/ICASSP48485.2024.10446751

[26] Vadim Popov et al., "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," arXiv: 2105.06337v2 [cs. LG] 5 Aug 2021 https://doi.org/10.48550/arXiv.2105.06337

[27] Kong, J. et al., "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, virtual, 2020. https://doi.org/10.48550/arXiv.2010.05646

[28] W. Han, Z. Zhang, Y. Zhang, et al., "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," arXiv: 2005.03191, 2020. https://doi.org/10.48550/arXiv.2005.03191

[29] Kriman, S. et al., "Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6124–6128, https://doi.org/10.48550/arXiv.1910.10261

[30] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, et al., "Nemo: a toolkit for building ai applications using neural modules," arXiv: 1909.09577, 2019. https://doi.org/10.48550/arXiv.1909.09577

[31] S. Chen et al., "Wavlm: Large-scale selfsupervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022. https://doi.org/10.48550/arXiv.2110.13900

[32] S. Ramoji, S. Ganapathy, "Supervised I-vector Modeling for Language and Accent Recognition," Computer Speech & Language (2019). https://doi.org/10.1016/j.csl.2019.101030

[33] Arun Babu et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in Proc. Of Interspeech, 2022, pp. 2278–2282. https://doi.org/10.48550/arXiv.2111.09296