Speech Embedding with Segregation of Para-linguistic Information for Tamil Language: a survey

Anosha Ignatius
dept. Computer Science and Engineering
University of Moratuwa
Sri Lanka
anoshai@uom.lk

Uthayasanker Thayasivam
dept. Computer Science and Engineering
University of Moratuwa
Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Abstract*—Deep Neural Networks (DNN) based speech embedding techniques have shown significant performance in speech processing applications. However, the presence of para-linguistic information such as speaker characteristics, accent, pronunciation, and emotion expression causes performance degradation in speech recognition where only the linguistic content is needed. Over the years many techniques have been proposed to address this problem by disentangling the para-linguistic content from the speech signal to learn better representations. The most common approach is to provide speaker level information at the input of acoustic model. In the case of low resource conditions when only a limited amount of transcribed speech data is available, unsupervised speech representation learning approach is adopted. A detailed study of research work related to disentangled speech representations is presented in this paper. Speech embedding techniques for Tamil Language are also discussed.**

***Index Terms*—speech processing, linguistic, para-linguistic information, speaker characteristics**

I. Introduction

Speech is a time-varying signal carrying multiple layers of information: linguistic information and para-linguistic information. Linguistic information refers to the meaning of the words being spoken while para-linguistic information refers to the residual information remaining after removing the verbal content from speech. Para-linguistic information covers the manner of speaking varied with the accent, pronunciation, emotional state and speaker traits. Speech processing applications that are focused only on the linguistic content such as Automatic Speech Recognition (ASR) and Speech Intent Recognition have yielded significant performance over the time. However, their performance can be greatly compromised due to mismatch between testing and training conditions. It is because they are subjected to variations in the manner of speaking caused by speaker characteristics, emotional expressions and environmental differences. Therefore, carrying out an extensive research on this area is of great importance for developing techniques that compensate for the speaker variability and improve the robustness of the systems. Several studies have investigated DNN based speech embedding models for removing the para-linguistic information from the speech signal while retaining the linguistic content.

DNN based speech recognition model is typically trained using acoustic feature vectors representing both the time domain and frequency domain information in the speech signal. Acoustic feature vectors are extracted using various feature extraction techniques such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC) and Perceptual Linear Prediction (PLP) [1]. MFCC features are most commonly used as input features in speech processing applications. Existing solutions for normalizing the variations caused by para-linguistic information incorporate speaker level information by augmenting the acoustic features with speaker representing vectors. This helps in disentangling the underlying speaker information in the speech signal by learning better representations. Vectors representing the speaker-specific information are obtained using a separate model. It involves mapping the variable length speech utterance to a fixed dimensional vector that captures speaker features. Prior studies have explored mainly two types of such techniques: i-vector approach [2] which is the conventional one and DNN based speaker embedding techniques. By providing speaker representations to the DNN as additional input features, DNNs are trained to be aware of the presence of speaker information. This approach is referred to as speaker aware training. Feature augmentation enables the acoustic model to normalise speaker effects, thus leading to a better generalization of the model to unseen conditions.

Under low resource conditions where a significant amount of labeled training data is not available, it is difficult to separate linguistically irrelevant speaker information encoded in the speech features. Supervised acoustic modelling relies on large amount of transcribed speech data to ensure the robustness against speaker variability. Thus, in the case of low resource scenario, unsupervised speech representations that separate speaker traits from linguistic content are desired to address this problem. Such representations learnt with an available large data set of unlabeled speech recordings can be used to learn an acoustic model with only a small amount of labeled training data. Tamil language comes under the low resource category and only a few studies have explored the problem of speaker variations.

The rest of the paper is organized as follows: Section II gives a detailed account of speaker embedding techniques, and discusses various methods for augmenting acoustic features with speaker vectors. Section III explores speech representations obtained using unsupervised learning approach and Section IV discusses Tamil speech recognition. The paper is concluded in Section V, with discussions in this research area.

II. Speaker Aware Training

ASR DNNs can be made invariant to speaker variability by providing speaker information to the DNN using speaker representing vectors. Augmenting the acoustic features with speaker vectors transforms the input features to a speaker normalised space.

A. Speaker Embedding

Speaker embedding encodes long term speaker characteristics that are difficult to learn with acoustic models using short term features. i-vector is an effective method to map a variable length speech segment to a low dimensional representation that captures speaker features. i-vector extraction involves Universal Background Model (UBM) and Total variability matrix. UBM, a speaker independent Gaussian Mixture Model (GMM) identifies and clusters similar acoustic features together. The total variability matrix represents the basis of the total variability space which models both the speaker variability and channel variability subspaces in a common low dimensional space. UBM and the total variability matrix are learnt in an unsupervised manner using expectation maximization algorithm. High-dimensional statistics from the UBM are then mapped into a low-dimensional i-vector. i-vectors have been extensively used in speaker identification and verification tasks. i-vectors are widely used for speaker normalization in ASR systems as well.

Speaker representation vectors can also be extracted using DNN based embedding techniques. In DNN based speaker embedding, a DNN which is trained to discriminate between speakers is used to extract the speaker representing vectors at the bottleneck layer. Recent DNN embeddings such as x-vectors extracted from a Time Delay Neural Network (TDNN) with a pooling layer [3] and h-vectors extracted using a hierarchical attention network [4] achieved better performance in speaker recognition tasks.

B. Speaker Normalisation

Speaker normalisation using feature augmentation techniques involves incorporating the speaker level information extracted as fixed dimensional vector in the input. It helps the DNN learn to normalize the speaker effects, thereby improving the performance. Several forms of feature augmentation include directly appending the speaker vectors to acoustic features and providing transformed features to the DNN during training.

The conventional i-vector based augmentation method uses i-vectors as additional input features. For each frame of a given utterance, same utterance level i-vector is concatenated with the acoustic features. Providing speaker information at the input enables the DNN to normalize the signal and make it invariant to speaker effects [5], [6]. Sri Garimella et al. proposed passing the i-vectors through a nonlinear hidden layer before combining them with the acoustic features [7]. In this model, connectivity from to the rest of network i-vectors is restricted to improve the robustness of the model. Using a nonlinear hidden layer to transform the i-vectors showed improvement over directly appending the to the acoustic features.

Two types of feature mapping neural networks: ivecNN and adaptNN are presented in the paper by Miao et al. where i-vectors are used as additional inputs to project acoustic features into a speaker normalized space [8]. In ivecNN, a bias vector is estimated and added to the original features making the resulting feature space speaker independent. The network takes i-vectors as the output and generates a feature shift as the output. The output is added to the acoustic features and fed to the DNN acoustic model. In adaptNN, multiple adaptation layers are used under the initial DNN acoustic model where each adaptation layer except the last one appends the i-vector to its output. By incorporating the i-vectors, the adaptation layers convert the original DNN input into more speaker independent features. Experimental results showed that these two networks achieve better performance over the original DNN with acoustic features.

Xiangang Li et al. investigated the idea of augmenting acoustic features with d-vectors, speaker embedding learnt using long short-term memory (LSTM) recurrent neural networks (RNNs) [9]. They proposed cross-LSTMP to conduct speaker recognition and speech recognition simultaneously. cross-LSTMP consists of two LSTMs, senone-LSTM for classifying senones and speaker-LSTM for classifying speakers. In each frame, previous activations of speaker-LSTM are fed into the senone-LSTM while the previous activations of senone-LSTM are into the speaker-LSTM to make the network speaker aware. Conducted experiments indicated that the proposed approach can effectively improve the performance by compensating for the speaker invariability in ASR.

Xiaodong Cui et al. proposed an embedding based speaker adaptive training approach [10] where speaker vectors are mapped to layer dependent element-wise affine transformations through a control network. Resulting affine transformations are applied to the internal feature representations at the outputs of the selected hidden layers in the acoustic model to facilitate speaker normalization. This approach outperformed feature augmentation with i-vectors.

Speaker Aware Speech Transformer, a standard speech transformer [11] with a speaker attention module (SAM) is presented in the paper by Zhiyun Fan et al. [12]. SAM includes a speaker knowledge block that consists of a group of i-vectors extracted from the training data and for each frame SAM generates a soft speaker embedding. As there are similarities among different speakers, speaker vector can be represented as a linear combination of a set of basic speaker representations. Thus, given a speech utterance, similarity of the acoustic feature vector and each i-vector from the group of basic speakers in speaker knowledge block is computed with the attention mechanism to obtain the weight for each basic i-vector. The soft speaker embedding is extracted as the weighted sum of the basic i-vectors and a weighted combined speaker embedding vector is fed to decoder This helps the model to normalize the speaker variations and leads to better generalization to unseen test speakers. Similar approach where attention mechanism is used to select the relevant speaker i-vectors for each frame from the memory is proposed by Jia Pan et al. [13] Speaker aware speech transformer proved to be more effective than other feature augmentation techniques using i-vectors.

III. Disentangled Speech Representations with unsupervised learning

Unsupervised learning method uses large amount unlabeled data to learn useful speech representations that can be incorporated into several downstream applications under low resource conditions. Auto encoders is one such network that involves the reconstruction of its inputs. Auto encoders consists of an encoding network that extracts a latent representation and a decoding network that tries to reconstruct the original data using the latent representation. By applying constraints, the network is made to learn a latent representation that discards irrelevant para-linguistic information while preserving the information necessary for perfect reconstruction.

Factorized Hierarchical Variational Auto encoder (FHVAE) [14] is proposed by Wei-Ning Hsu et al. to learn disentangled representations from sequential data. It can be used to separate linguistic content and speaker information in speech signal in an unsupervised way. The FHVAE learns to factorize sequence-level and segment-level attributes of speech into different latent variables [15], [16]. Speaker identity affects the speech features at sequence-level while phonetic content affects the speech features at utterance level. Hence, sequence-level attributes show a small amount of variation within an utterance and segment-level attributes show similar amounts of variation within and across utterances. Based on this, different sets of latent variables are generated. Latent variable that encodes the linguistic content while discarding the speaker characteristics is used in robust speech recognition task.

Jan Chorowski et al. presented a Vector Quantized Variational Auto encoder (VQ-VAE) that learns a representation to capture high level semantic content while being invariant to the underlying speaker information in the signal [17]. It conditions the decoder on speaker identity which frees up the encoder from having to capture speaker features, thus resulting in a more speaker invariant representation. Best representations are obtained when MFCC features are used as inputs and raw waveforms are used as targets. Waveforms are reconstructed using a WaveNet decoder that combines auto-regressive information about past wave samples, speaker information and latent information extracted by the encoder. VQ-VAE yielded a significant performance in the phonetic unit discovery task indicating a better separation between phonetic content and speaker information.

Another unsupervised objective, Auto-regressive Predictive Coding (APC) [18]–[20] is proposed by Yu-An Chung et al. can be used as a pre-training approach for learning meaningful and transferable speech representations. It is trained to predict the spectrum of future frames n steps ahead of the current one using the past values to infer more global structures in speech rather than exploiting local smoothness of the speech signal. Unlike the previously discussed research studies which attempted to discard the irrelevant information, APC model aims to preserve as much information such that the extracted representations could be used for a variety of downstream tasks. Experiments conducted using intermediate representations obtained from APC model on different speech applications such as speaker verification, phone classification and automatic speech recognition indicate that different levels of speech information are captured by the APC model at different layers. Lower layers contain more speaker information while the upper layers provide more phonetic content. Combination of the internal representations extracted across different layers could be beneficial in learning disentangled representations for speech recognition tasks.

The aforementioned research studies show that unsupervised learning method can be effectively applied on large amounts of unlabeled data to extract speaker invariant features that helps in improving the robustness of low resource speech recognition systems.

IV. Tamil Speech Recognition

Improving speech recognition in tamil language is an active research field and it is a challenging task due to the unavailability of a large corpus of tamil speech. Though, many research studies have attempted to build a reasonably good speech recognition systems [21], [22]. Few studies have investigated the variations caused by speaker characteristics. Akila et al. addressed the effects of speech rate variability by applying time normalization to the speech signal [23]. The speech signal is categorized as slow, normal and fast speech using features such as the sound intensity level and time duration. If the speech rate is either slow or fast, time normalisation is applied. In the paper by Madhavaraj et al. speaker adaptive training is used with Maximum likelihood linear transformation to improve the recognition [23].

V. Summary

The study presented an overview of various techniques that suppress the effects of para-linguistic information in speech recognition systems. Disentangling the linguistically irrelevant information is important to alleviate the problem of mismatch between training and testing conditions. Existing solutions were studied under two categories: speaker embedding based feature augmentation and unsupervised speech representation learning. Incorporating the speaker information into the acoustic model through speaker embedding have resulted in improved recognition rate but such supervised acoustic models require large amount of labeled training data for a more robust system. In case of low resource scenario, unsupervised learning method can be adopted where speech representations learnt from a large-scale unlabeled data are used for downstream tasks with a limited amount of labeled training data. This is proved to be a powerful approach to achieve better generalisation for acoustic models. Speech Recognition systems for Tamil language which lacks speech data could exploit these methods to achieve better performance.

Acknowledgment

This research was supported by Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education funded by the World Bank.

References

[1] V. Z. Këpuska and H. A. Elharati, 'Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions,' J. Comput. Commun., vol. 03, no. 06, pp. 1-9, 2015, doi: 10.4236/jcc.2015.36001.

[2] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," Odyssey 2010 Speak. Lang. Recognit. Work., no. May 2014, pp. 76–82, 2010.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 5329–5333, 2018, doi: 10.1109/ICASSP.2018.8461375.

[4] Y. Shi, Q. Huang and T. Hain, "H-Vectors: Utterance-Level Speaker Embedding Using a Hierarchical Attention Model," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7579-7583, doi: 10.1109/ICASSP40776.2020.9054448.

[5] I. L.-M. Andrew Senior, "IMPROVING DNN SPEAKER INDEPENDENCE WITH I -VECTOR INPUTS," 2014 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 225–229, 2014.

[6] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. H. L. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2015-January, pp. 2854–2857, 2015.

[7] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi, "Robust i-vector based adaptation of DNN acoustic model for speech recognition," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2015-Janua, pp. 2877–2881, 2015.

[8] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. September, pp. 2189–2193, 2014.

[9] X. Li and X. Wu, "Modeling speaker variability using long short-term memory networks for speech recognition," in Interspeech, 2015.

[10] Xiaodong Cui, Vaibhava Goel, and George Saon, "Embedding-Based Speaker Adaptive Training of Deep Neural Networks," in Interspeech 2017. 122-126. 10.21437/Interspeech.2017-460.

[11] Xiaorui Wang Yuanyuan zhao, Jie Li and Yan Li, "The speechtransformer for large-scale mandarin chinese speech recognition," in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[12] Fan, Z., Li, J., Zhou, S., and Xu, B., "Speaker-Aware Speech-Transformer," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 222-229.

[13] Jia Pan, Diyuan Liu, Genshun Wan, Jun Du, Qingfeng Liu, and Zhongfu Ye, "Online speaker adaptation for lvcsr based on attention mechanism," in Proceedings, APSIPA Annual Summit and Conference, 2018, vol. 2018, pp. 12–15.

[14] W. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in Proc. NIPS, 2017, pp. 1876–1887.

[15] W. Hsu and J. R. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in Proc. ICASSP, 2018, pp. 5614–5618.

[16] Siyuan Feng and Tan Lee, "Improving Unsupervised Subword Modeling via Disentangled Speech Representation Learning and Transformation," in Interspeech 2019.

[17] J. Chorowski, R. J. Weiss, S. Bengio and A. van den Oord, "Unsupervised Speech Representation Learning Using WaveNet Autoencoders," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2041-2053, Dec. 2019, doi: 10.1109/TASLP.2019.2938863.

[18] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An unsupervised autoregressive model for speech representation learning," in Interspeech, 2019.

[19] Y. Chung and J. Glass, ”Generative Pre-Training for Speech with Autoregressive Predictive Coding,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 3497-3501, doi: 10.1109/ICASSP40776.2020.9054438.

[20] Yu-An Chung, Hao Tang and James Glass, ”Vector-Quantized Autoregressive Predictive Coding,” in Interspeech, 2020.

[21] I. Kalith, David Asirvatham, and Raisal Ismail, “ Context-dependent Syllable Modeling of Sentence-based Semi-continuous Speech Recognition for the Tamil Language,” 2017. Information Technology Journal. 10.3923/itj.2017.

[22] Lokesh, S., Kumar, P.M., Devi, M., Panchatcharam, P., and Babu, G.C. (2018). ”An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self-Organizing Map,” Neural Computing and Applications, 31, 1521-1531.

[23] Akila, A. and Chandra, E. ”Performance enhancement of syllable based Tamil speech recognition system using time normalization and rate of speech,” CSIT 2, 77–84 (2014). doi: 10.1007/s40012-014-0044-6.

[24] A. Madhavaraj and A. G. Ramakrishnan, ”Design and development of a large vocabulary, continuous speech recognition system for Tamil,” 2017 14th IEEE India Council International Conference (INDICON), Roorkee, 2017, pp. 1-5, doi: 10.1109/INDICON.2017.8488025.