What does an End-to-End Dialect Identification Model Learn about Non-dialectal Information?

Shammur A Chowdhury[1], *Ahmed Ali*[1], *Suwon Shon*[2], *James Glass*[3]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]ASAPP Inc., New York, NY, USA
[3]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

{shchowdhury,amali}@hbku.edu.qa, sshon@asapp.com, glass@mit.edu

Abstract

An end-to-end dialect identification system generates the likelihood of each dialect, given a speech utterance. The performance relies on its capabilities to discriminate the acoustic properties between the different dialects, even though the input signal contains non-dialectal information such as speaker and channel. In this work, we study how non-dialectal information are encoded inside the end-to-end dialect identification model. We design several proxy tasks to understand the model's ability to represent speech input for differentiating non-dialectal information – such as (a) gender and voice identity of speakers, (b) languages, (c) channel (recording and transmission) quality – and compare with dialectal information (i.e., predicting geographic region of the dialects). By analyzing non-dialectal representations from layers of an end-to-end Arabic dialect identification (ADI) model, we observe that the model retains gender and channel information throughout the network while learning a speaker-invariant representation. Our findings also suggest that the CNN layers of the end-to-end model mirror feature extractors capturing voice-specific information, while the fully-connected layers encode more dialectal information.

Index Terms: dialect identification, speaker information, language identification, end-to-end model, interpretability

1. Introduction

The end-to-end deep neural network for speech technologies such as automatic speech recognition (ASR) [1, 2, 3, 4], dialect, language and speaker identification [5, 6, 7, 8, 9, 10, 11], provides a simplified and flexible training architecture with improved performance. However, this engineering flexibility comes at the expense of model interpretability, along with blankness and abstraction regarding different encoded information and representations in the layers of the model.

To interpret the encoded information, many studies have explored the phonetic [12, 13, 14], graphemic [12, 15] along with different articulatory information in end-to-end ASR models using clustering and classification techniques as downstream tasks. Apart from phonetic information, speaker properties such as style and accent [16], are also analyzed to understand the intermediate layer representation. Studies like [17, 18] correlated the behaviors of RNN gates with phoneme boundaries while others clustered the neurons of an end-to-end ASR [19] system. The work in [20] visualize skip connections in speech enhancement models. Phonetic properties are also investigated for speaker embedding models [21, 22]. However, no significant effort has been given to understand end-to-end language/dialect identification models for the encoded information.

In this study, we investigate an *end-to-end **A**rabic **D**ialect **I**dentification* (**ADI**) model for *encoded dialectal* and *non-dialectal* information. We explore the learned internal representations for *gender*, *voice identity*, *language* and *channel-based characteristics* along with *dialectal properties*.

Dialect identification is a special case of the language identification (LID) task and is relatively unexplored compared to language and speaker recognition. The task of identifying dialects from the speech signal is extremely challenging since the performance depends on the ability of the model to discriminate the acoustic dissimilarities between dialects within the same language family. *Arabic* is an appropriate language for the task due to its uniqueness as a shared language with 22 countries and having more than 20[1] mutually incomprehensible dialects, with a common phonetic and morphological inventory.

For the study, we exploit the layer-wise embeddings to observe what dialectal and non-dialectal information the network is encoding. We designed various proxy tasks like classification and verification. Since *speaker*, *language*, *channel information* – signal recording and transmission quality, of the speech impose different characteristics on the acoustic signal, the network is acquainted to this information along with dialectal properties. Hence, we probe the network to observe if such non-dialectal information (speaker, language and channel) are captured by the ADI network.

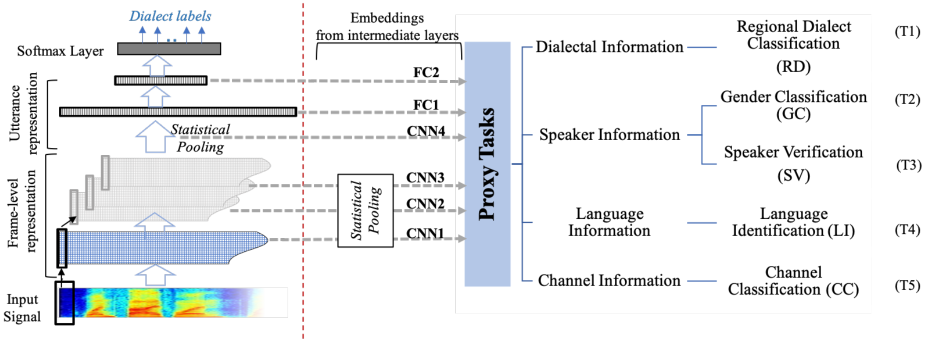
Our contributions are: (*i*) understanding what non-dialectal information, in terms of speaker, language and channel, is encoded in the ADI network; (*ii*) exploring which layers capture more dialectal information than others. To the best of our knowledge, this is the first attempt to understand non-task oriented information encoded in a dialect (or any variation of language) identification model.

The structure of this paper is as follows. We present a detailed description of the methodology followed in this study along with datasets and architectures in Section 2. Following in Section 3, we report and discuss the findings of the study. Finally, we conclude our work in Section 4.

2. Experimental Methodology

In Figure 1, we present the architecture of the ADI model and the experimental flow of the study. For the task, we exploited an end-to-end ADI architecture (see Section 2.1) proposed in [10]. We extracted the utterance level representation (embedding) from each layer of the ADI network. Afterwards, we adapted different probing techniques, discussed in Section 2.2. We trained separate classifiers for tasks T1, T2, T4, T5 (as in Figure 1) using embeddings from each layer. As for T3, we evaluated the embeddings using verification pairs for speaker

[1] In this study, we explore an ADI model trained with 17 dialects.



(a) Original Dialect Identification model

(b) Proxy task list

Figure 1: *The experimental flow of the study. Figure 1(a) presents the architecture of the Arabic end-to-end dialect identification (ADI) system used for the analysis. Figure 1(b) presents the steps used to extract the embeddings from different layers, following the several proxy tasks used to probe the model for the encoded dialectal and non-dialectal information. T[1-5] are the task ids corresponding to the explored information. FC - fully-connected layer, CNN - convolution layer.*

identity (details in Section 2.2-T3).

2.1. E2E Dialect Identification Model

We adopted the end-to-end system architecture proposed in [10], trained using the 'Arabic Dialect Identification 17' (ADI17) dataset [23, 24], referred to as the ADI-17 model.[2]

As input to the model, we extracted a total of 40 coefficient MFCCs features from a spectrogram computed with a 25ms window and 10ms frame-rate from 16kHz audio. The architecture of the model includes four temporal convolution neural networks (1D-CNNs), followed by a global (statistical) pooling layer to aggregates the frame-level representations to utterance level representations.[3] For the CNN layers, we used filter sizes of 40×5, 1000×7, 1000×1, 1000×1 with 1-2-1-1 strides and 1000-1000-1000-1500 filters respectively. This utterance level representation is then passed to two fully connected layers (hidden units: 1500 and 600). We used Rectified Linear Units (ReLUs) as activation functions of the network.

To train the network, the stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 is used. The overall performance of the trained ADI-17 model using official MGB-5 dialect test set [24] are: accuracy - 82.0% and F_1 - 82.7%.

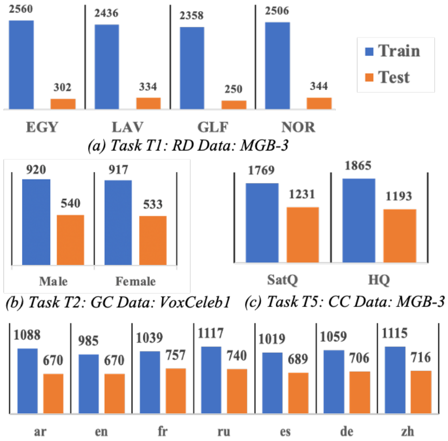
2.2. Proxy Tasks

Given the extracted embeddings from each of 6 layers of ADI-17 model, we designed 5 proxy tasks, (total 6 × 4 = 24 classification tasks and 6 × 1 speaker verification task per language), to examine the ability of the system for encoding dialectal and non-dialectal information. We considered the following tasks for our study:

T1 - Regional Dialect Classification (RD)

For regional dialect classification, we designed a simple feed-forward neural network with a hidden layer of size 500 and a softmax output layer. The input of the network is the embedding extracted from each intermediate layer of the ADI-17 model. We trained the network for 100 epochs with a batch size of 128 and SGD optimizer with a learning rate of 0.01.

For training the model, we used the Arabic ADI-5 dataset [25], which is composed of the following five dialects: Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African Region (NOR) and Modern Standard Arabic[4] (MSA). The dataset contains satellite cable recording (SatQ) in the official training split and high-quality (HQ) broadcasts videos for development and test set. For the classification, we used the balanced train set to design the proxy task and tested using the test split. Details of the class distribution is reported in Figure 2(a).



(a) Task T1: RD Data: MGB-3

(b) Task T2: GC Data: VoxCeleb1 *(c) Task T5: CC Data: MGB-3*

(d) Task T4: LI Data: Common Voice

Figure 2: *Data distribution for Proxy classification tasks.*

T2 - Gender Classification (GC)

For gender classification, we trained a single layer feed-forward network (250 hidden units) with a softmax output layer, for 20 epochs using a batch size of 128 and SGD optimizer with a learning rate of 0.01.

For the task, we trained proxy classifiers using *VoxCleleb1*-test [8] (English) dataset. The VoxCeleb1 is a gender balanced dataset that includes videos of celebrities, from different ethnicities, accents, professions and ages, uploaded to YouTube. Detailed label distribution[5] for the task is given in Figure 2(b).

T3 - Speaker Verification (SV)

For voice identity verification, we performed a 'generic speaker verification' using pairs of input signals and verifying if they are

[2] `https://github.com/swshon/arabic-dialect-identification`

[3] We followed similar approach to extract utterance level representation from the first 3 CNN layers for our study (see Figure 1).

[4] For the classification, we ignored the instances labelled as MSA.

[5] with no overlapping speakers in train-test

from the same speaker or not. We extracted length normalized embeddings from each layer of the ADI-17 model and computed the cosine similarity between pairs. We constructed these verification pair trials by randomly picking up utterance pairs from speakers with the same gender, maintaining a balanced distribution between positive and negative targets.

We used a multi-lingual subset of the *Common Voice* dataset [26] and the *Voxceleb1* official verification test set (Voxceleb1-tst)[6] [8]. The *Common Voice corpus* contains more than $2,500$ hours of speech data from $\approx$ 39 languages[7], collected and validated via a crowdsourcing approach. This is one of the largest multilingual datasets available for speech research, recorded using a website or an iPhone application available from the Common Voice project.

We performed speaker verification using three out-of-domain languages including English (en) – from the Voxceleb1-tst, and a subset[8] from the Common Voice – Russian (ru) and Chinese (zh) datasets. Details of the the verification pairs is given in Table 2.

For performance comparison, we also evaluate these datasets using a task-specific model designed to recognize speakers. The **S**peaker **R**ecognition (**SR**) model, adapted from [22], is trained using the Voxceleb1 development set (containing 1211 speakers and $\approx$ 147K utterances), using the same architecture and the parameters mentioned in Section 2.1. We then performed speaker verification, using the embedding from the last intermediate layer (second fully-connected layer, FC2) of the SR model.

T4 - Language Identification (LI)

For the language identification task, we designed classifiers (using a similar architecture mentioned in T2:GC) for discriminating between the 7 languages selected from the Common Voice dataset. The language subset used for this study includes – Arabic (ar), English (en), Spanish (es), German (de), French (fr), Russian (ru) and Chinese (zh). The distribution of the datasets for training and testing the classifiers are shown in Figure 2(d).

T5 - Channel Classification (CC)

For measuring the ability of the ADI-17 to capture information regarding transmission and signal recording quality, we designed binary classifiers from the network layers. Using a similar architecture as mentioned in T2:GC, the classifier output labels indicating the input signal quality as Satellite recording (SatQ) *vs* High quality archived videos (HQ).

For this task, we combined ADI-5 train (includes SatQ), dev (HQ) and test (HQ) set and randomly picked balanced samples from each class. These selected samples are then divided into train-test using 60-40% split for the experiment. Distribution of the dataset is given in Figure 2(c).

2.3. Evaluation Measures

To asses the performance of classification tasks (T1:RD, T2:GC, T4:LI and T5:CC), we reported macro F-measure – where the result is calculated by averaging the performance on each label. As for the speaker verification (T3:SV), we report Equal Error Rate (EER) – measuring the value at which the false-reject (miss) rate equals the false-accept (false-alarm) rate.

3. Results and Discussions

In this Section, we report our findings for the proxy tasks mentioned in Section 2.2. The performance of the proxy tasks used to probe the intermediate layer representations of ADI-17 models are presented in Table 1 and Table 2.

T1 - Encoding Regional Dialect Information

For the regional dialect classification task, the performance (as shown in Table 1) across the intermediate layer indicates that the dialectal information is encoded in the fully-connected layers rather than in temporal CNNs. Similar performance is observed using the FC1 layer compared to the performance of $F_1 = 58.66\%$ from the output layer of ADI-17 model.

T2 - Encoding Speaker Gender Information

The performance of the gender classifier suggests that the model is encoding gender information of the speaker throughout the network.[9] The innateness of this information is reflected by the high performance of lower-level CNNs, as given in Table 1.

T3 - Encoding Voice Identity of Speakers

The EER for the speaker verification task is presented in Table 2. The homogeneity pattern of the task performance suggests that the ADI-17 network is not able to distinguish vocal information across the different layers. However, a slight improvement is observed only after the CNN4 layer for all the test data. This observation can be hypothesised as the network capturing language features instead of speaker information.

To gain better insight of the ideal performance expected from a model that captures speaker voice identity, we simultaneously reported results for the test sets using a speaker recognition (*SR*) model, trained using Voxceleb1-dev dataset with similar architecture as ADI-17 (model refereed as *SR* in Table2).

The overall result indicates that the trained *SR* has the ability to capture language-independent speaker information. Our finding suggests that the ADI model refrains from capturing speaker voice identity. The hypothesis is that the CNNs is capturing some vocal features which are then encoded in FC1 layer giving a slight improved EER. To verify, we designed our next proxy task as a language recognition problem (T4:LI).

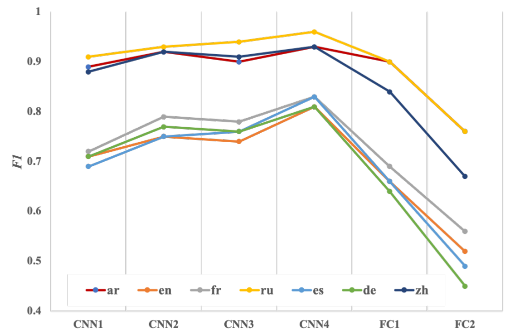


Figure 3: *Reported class-based F-measure for the language identification proxy task (T4:LI). The class-labes are languages including: ar-Arabic, en-English, fr-French, de-German, ru-Russian, es-Spanish and zh-Chinese.*

[6] In this case we used the official verification pairs to evaluate.

[7] last accessed: April 10, 2020

[8] Randomly selected ≈4 hours from each language

[9] A similar pattern is observed when experimented with ADI-5 data. For brevity, we are not reporting the results of the experiment.

Table 1: *Reported macro F-measure for proxy tasks T1, T2, T4 and T5. *For the classification, we ignored the instances labelled as MSA from the data, hence using 4 labels of dialects rather than 5.*

Proxy Tasks	Dataset Used	CNN1	CNN2	CNN3	CNN4	FC1	FC2	#. class labels
T1: RD	MGB-3	0.12	0.12	0.12	0.20	**0.58**	0.55	4*
T2: GC	Voxceleb1-tst	0.96	0.97	0.98	**0.99**	**0.99**	**0.99**	2
T4: LI	Common Voice	0.79	0.83	0.83	**0.87**	0.76	0.60	7
T5: CC	MGB-3	0.88	0.90	0.89	**0.92**	0.90	0.88	2

Table 2: *Reported EER for proxy task T3: Speaker Verification (SV). Positive (+) represents percentage of pairs from the same speakers. Both positive and negative pairs are extracted from same gender speakers. Lang. is Language of the dataset.*

Models →			*ADI-17*						*SR*
Dataset Used	**Speakers (+)**	**Lang.**	**CNN1**	**CNN2**	**CNN3**	**CNN4**	**FC1**	**FC2**	**FC2**
Voxceleb1-tst	40 (50%)	en	29.89	27.47	27.91	26.74	22.27	26.02	**6.81**
Common Voice	28 (55.3%)	ru	18.64	18.56	17.58	17.40	13.47	17.04	**4.05**
	69 (48.9%)	zh	16.17	15.47	14.39	14.04	13.55	15.63	**5.47**

T4 - Encoding Language Information

To access the ability of the ADI-17 model to encode vocal features capturing language information, we designed multi-class classifiers with 7 languages as class-labels. The overall performance of the proxy models are reported in Table 1. Our result suggests that the utterance-level embedding from CNN4 captures the language representation better for the classification task yielding the best performance. Further analysis of class wise performance (in Figure 3) showed a similar pattern with CNN4 giving best F_1 for all the classes. Moreover, when analyzing the confusion matrix, we observe that the proxy classifier successfully discriminates between Russian, Arabic and Chinese, while a confusion of $\approx 12 - 14\%$ is shown between the languages – English, German, French and Spanish. This observation aligns with our hypothesis in *T3:SV*, suggesting that the CNNs act as feature extractors of vocal tract information that provide a better representation for the language identification task.

T5 - Encoding Channel Information

The performance of input channel classification is reported in Table 1. These high and homogeneous performances of the proxy classification tasks suggest that channel information is highly embedded across the layers of ADI-17 model. Thus influencing the network performance based on the channel quality of the input signal. Implicating the importance of training dialect identification models with datasets from different channels for model generalization.

Key Observations

Using utterance-level embeddings, we observed that the representations from higher intermediate layers (FCs) contain significantly more dialectal information than lower CNN layers. This indicates the ability of the FC-layers to learn task-oriented information compared to the CNNs.

Probing the network for non-dialectal speaker information indicates that the model captures gender information throughout the network without any language dependency. However, the same model refrains from encoding knowledge to distinguish the voice identity of speakers.

When investigated for encoded language information, we observed that the embeddings from CNN layers (specifically CNN4) significantly outperforms all other network layers. We noticed a gradual decline in performance when FC layers' embedding is used. This observation is contrary to the pattern observed when probing the network for dialectal information. Thus indicating the importance of FCs for learning dialectal information and CNNs for capturing general vocal features that can be used to discriminate between the languages.

As for the network encoding channel information, we noticed that the classification performance is highest using CNN4 embedding and then decreases slowly in the succeeding layers, with a drop of almost 2-4% when the FC2 layers are reached. This indicates that the channel information is encoded in the model however is less representative in the FC2 layer.

The findings from T1:RD, T3:SV, T4:LI and T5:CC infer that the CNN layers are capturing vocal features and imitate a feature extractor of the acoustic model, whereas the FC layer acts a dialect classifier. Thus showing that the higher layers encode very task-specific features, which aligned with the findings of neural networks interpretability (e.g. [27]) studies.

4. Conclusion

In this study, we analyzed an end-to-end Arabic dialect identification system for both dialectal and non-dialectal encoded information. To investigate the intermediate representation, we adopted several proxy tasks using multi-lingual datasets.

From our experimental results, we observe that *speaker gender information is embedded* throughout the network. A *similar pattern* is observed for *channel information*. Unlike gender information, speaker verification tasks shows the network is *learning speaker-invariant representations*. From the language identification task, we observe that the CNN layers of the network performed significantly well, in contrast to the regional dialect classification task where FCs outperform the CNNs. This suggests that *CNNs are better in capturing vocal representation* thus performing better at non-dialectal tasks, whereas *FCs encode more dialectal information*.

To the best of our knowledge, this is the first attempt to investigate non-task oriented information in a dialect (or language) identification model. Some non-dialectal information such as 'channel' information can be discarded using better task design and data source variation. However, in future we plan to explore if having channel information can aid in model generalization. In addition, we also plan to extend our analysis to other E2E models.

5. References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[2] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[3] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6460–6464.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] M. Jin, Y. Song, I. V. McLoughlin, W. Guo, and L.-R. Dai, "End-to-end language identification using high-order utterance representation with bilinear pooling," 2017.

[6] T. N. Trong, V. Hautamäki, and K.-A. Lee, "Deep language: a comprehensive deep learning approach to end-to-end language recognition." in *Odyssey*, 2016, pp. 109–116.

[7] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[10] S. Shon, A. Ali, and J. Glass, "Convolutional neural network and language embeddings for end-to-end dialect recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 98–104.

[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)*. IEEE, 2018, pp. 5329–5333.

[12] Y. Belinkov, A. Ali, and J. Glass, "Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition," *Proc. Interspeech 2019*, pp. 81–85, 2019.

[13] L. Bai, P. Weber, P. Jancovic, and M. J. Russell, "Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features." in *Interspeech*, 2018, pp. 1472–1476.

[14] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "On the role of nonlinear transformations in deep neural network acoustic models." in *Interspeech*, 2016, pp. 803–807.

[15] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems*, 2017, pp. 2441–2451.

[16] Z. Elloumi, L. Besacier, O. Galibert, and B. Lecouteux, "Analyzing learned representations of a deep asr performance prediction model," in *Blackbox NLP Workshop and EMLP 2018*, 2018.

[17] Y.-H. Wang, C.-T. Chung, and H.-Y. Lee, "Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries," *Proc. Interspeech 2017*, pp. 3822–3826, 2017.

[18] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5140–5144.

[19] A. Krug, R. Knaebel, and S. Stober, "Neuron activation profiles for interpreting convolutional speech recognition models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*, 2018.

[20] J. F. Santos and T. H. Falk, "Investigating the effect of residual and highway connections in speech enhancement models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*, 2018.

[21] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Interspeech*, 2017, pp. 1497–1501.

[22] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1007–1013.

[23] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, "Adi17: A fine-grained arabic dialect identification dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8244–8248.

[24] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1026–1033.

[25] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic mgb-3," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 316–322.

[26] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[27] F. Yu, Z. Qin, and X. Chen, "Distilling critical paths in convolutional neural networks," *arXiv preprint arXiv:1811.02643*, 2018.