A FACTORIAL DEEP MARKOV MODEL FOR UNSUPERVISED DISENTANGLED REPRESENTATION LEARNING FROM SPEECH

Sameer Khurana[⋆]*, Shafiq Rayhan Joty*[†]*, Ahmed Ali*[♣]*, James Glass*[⋆]

[⋆] MIT Computer Science & Artificial Intelligence Laboratory, Cambridge, MA, USA
[†] School of Computer Science and Engineering, NTU, Singapore
[♣] Qatar Computing Research Institute, HBKU, Doha, Qatar

ABSTRACT

We present the **F**actorial **D**eep **M**arkov **M**odel (FDMM) for representation learning of speech. The FDMM learns disentangled, interpretable and lower dimensional latent representations from speech without supervision. We use a static and dynamic latent variable to exploit the fact that information in a speech signal evolves at different time scales. Latent representations learned by the FDMM outperform a baseline i-vector system on speaker verification and dialect identification while also reducing the error rate of a phone recognition system in a domain mismatch scenario.

Index Terms— Disentangled Representation Learning, Variational Inference, Factorial Deep Markov Model

1. INTRODUCTION

Our interest in unsupervised speech processing stems from the desire to depart from expert based, fully supervised automatic speech recognition systems to the decipher-based scenario [1] where unlabeled speech and non-parallel text are available. In this scenario, a machine would have to learn to read and listen from scratch without correspondences between speech and text. Unsupervised representation learning can be seen as tackling the listening part of the larger problem. Another motivating factor for our work is unsupervised spoken language acquisition – the problem of discovering discrete linguistic structure from speech. The problem of acoustic unit discovery (AUD) [2] falls under this category. The task is to cluster similar sounding acoustic segments, thereby discovering sound units that occur frequently in a speech corpus. A lower dimensional structured latent space can make the problem easier by reducing the number of parameters needed to build an AUD clustering model [3].

In this work, we propose a novel generative model, the factorial deep markov model (FDMM) (Section 4) that learns disentangled and intepretable representations from speech without supervision. At a high level, the FDMM is just a variational auto-encoder (VAE) [4] which, in addition to the usual encoder and decoder neural nets, has a transition neural net that models the *Markovian* dynamics in the latent space. The model is trained using Stochastic Variational Inference (SVI), an optimization-based approximate inference method (Section 3). We evaluate our model on speaker verification, dialect identification and domain mismatched ASR tasks (Section 5) and show that it successfully encodes content and style/domain information in two independent (in the prior) latent variables.

2. RELATION TO PRIOR WORK

We build on the excellent work of Hsu et. al [5] which introduced a factorial hierarchical VAE (FHVAE) for disentangled representation learning from speech. Like the FDMM, the FHVAE also has content and style/domain latent variables $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ respectively. To exploit the multi-scale information present in the speech signal, the FHVAE has a fixed sequence level prior, $\boldsymbol{\mu_2}$ on $\boldsymbol{z_2}$ to encourage $\boldsymbol{z_2}$ to evolve at a lower time resolution than $\boldsymbol{z_1}$. Estimates for $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ are given by neural network encoders mapping the observation space to the latent space, while a lookup table, $\mathcal{L}$, indexed by sequence ID provides estimates for $\boldsymbol{\mu_2}$, ensuring that $\boldsymbol{\mu_2}$ is sampled only once per sequence. As $\mathcal{L} \in \mathbb{R}^{N \times d}$, its size grows with the number of training sequences (N), which makes the FHVAE impractical to train on large datasets. In recent follow up work Hsu et. al [6] propose a training methodology based on hierarchical sampling of the training sequences that makes training an FHVAE possible on large datasets. Alternatively, in our work, we show that increasing the time resolution for $\boldsymbol{z_1}$, instead of decreasing it for $\boldsymbol{z_2}$ and capturing temporal dynamics of the speech signal in the model, allows us to eliminate $\boldsymbol{\mu_2}$ altogether. In this way, our model overcomes the shortcoming of the FHVAE and is the **key contribution** of our work.

Our model is directly inspired by Krishnan et. al [7], which introduces a Deep Markov Model and trains it using SVI. We extend their work and introduce a static random variable to encourage disentanglement in the latent space. A parallel work [8] introduced a disentangled sequential VAE (DSVAE), which, like us, models $\boldsymbol{z_1}$ at a higher time-resolution than $\boldsymbol{z_2}$. Unlike the DSVAE, we model state transition probabilities in the prior and also perform a sys-

tematic evaluation of different posterior inference networks (Section 4.2, Table 1).

3. VARIATIONAL INFERENCE

Approximate inference techniques can be categorized into sampling based methods such as *Markov Chain Monte Carlo* and *Variational Methods* [9]. In this paper we use a Variational Method which is discussed in some detail below. For sampling-based methods readers are referred to some excellent work presented in [10, 11].

Variational inference turns the problem of inference into optimization. In Variational inference we approximate the intractable posterior distribution p with a simpler distribution $\boldsymbol{q}$, parameterized by $\boldsymbol{\phi}$. Different values of $\boldsymbol{\phi}$ denote different members of the family $\boldsymbol{q}$; $\boldsymbol{q}$ is called the *variational family* and $\boldsymbol{\phi}$ are the *variational free parameters*. The optimization objective is then to find the member of the family $\boldsymbol{q}$ that is closest to the true posterior $\boldsymbol{p}$. Closeness between the two distributions is measured using the KullbackLeibler (KL) divergence between the two distributions [9]. Formally, we can write the optimization objective in terms of KL:

$$\boldsymbol{q^\star(h)} = \arg\max_{\boldsymbol{q(h)}} \mathrm{KL}(\boldsymbol{q(h)} \,||\, \boldsymbol{p(h|x)}) \tag{1}$$

which is equivalent to writing in terms of the free parameter $\boldsymbol{\phi}$:

$$\boldsymbol{\phi^\star} = \arg\max_{\boldsymbol{\phi}} \mathrm{KL}(\boldsymbol{q_\phi(h)} \,||\, \boldsymbol{p(h|x)}) \tag{2}$$

where, $\boldsymbol{h}$ is the latent space and $\boldsymbol{x}$ the observation space.

Using the formula for $\mathrm{KL}(\boldsymbol{q}||\boldsymbol{p}) = \sum_h \boldsymbol{q(h)}\ \boldsymbol{log}\frac{\boldsymbol{q(h)}}{\boldsymbol{p(h|x)}}$ and the fact that $\mathrm{KL} > 0$, it is straightforward to show that minimizing KL is equivalent to maximizing the lower bound on the model likelihood $\boldsymbol{p(x)}$, also known as the model evidence. This objective function is popularly know as the Evidence Lower Bound or ELBO [4] and is given by:

$$\boldsymbol{\mathcal{L}(\phi,\theta)} = -\boldsymbol{E_q}\left[\boldsymbol{log}\ \boldsymbol{p_\theta(x|h)}\right] + \mathrm{KL}(\boldsymbol{q_\phi(h)}||\boldsymbol{p(h)}) \tag{3}$$

$\boldsymbol{\mathcal{L}}$ is maxed when the reconstruction loss, $-\boldsymbol{E_q}\left[\boldsymbol{log}\ \boldsymbol{p(x|h)}\right]$ and the KL term is minimized. The KL term acts as a regularization that encourages $\boldsymbol{q}$ to be diverse [9].

4. THE FACTORIAL DEEP MARKOV MODEL

4.1. FDMM Description

The generative model has two random variables: a) segment level, $\boldsymbol{y}$ and (b) frame level $\boldsymbol{z}$. The difference in time resolution encourages disentanglement in the latent space such that $\boldsymbol{z}$ encodes the content information while $\boldsymbol{y}$ encodes style/domain information. More formally, the generative model is defined as :

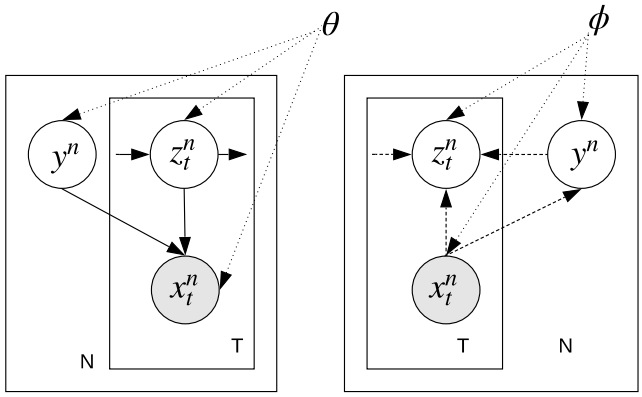

Fig. 1. Proposed generative model on the left and the corresponding inference model on the right. $\boldsymbol{z_t}$ and $\boldsymbol{y}$ are latent random variables, $\boldsymbol{x_t}$ is the observed random variable, T is the number of frames in the acoustic segment, $\boldsymbol{x_{1:T}}$. Model and posterior parameters are θ and ϕ respectively. $\boldsymbol{y}$ and $\boldsymbol{z}$ encodes information present at different time scales in the speech signal

$$p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}) = \prod_{n=1}^{N} \boldsymbol{p_\theta(y^n)} \prod_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{p_\theta(z_t^n|z_{t-1}^n)} \boldsymbol{p_\theta(x_t^n|z_t^n,y^n)} \tag{4}$$

where, $\boldsymbol{x_t^n}$ is the t^{th} acoustic frame in an acoustic segment of length T belonging to the n^{th} sequence in a dataset containing a total of N sequences, $\boldsymbol{y^n}$ is the segment level random variable. The above factorization can be read off from the graph structure given in Figure 1.

Each of the conditional probability distribution on the right hand side of the equation 4 is given as follows:

$$\boldsymbol{p_\theta(x|y,z_t)} = \boldsymbol{\mathcal{N}(f_\theta^\mu(y,z_t), f_\theta^{\sigma^2}(y,z_t))} \tag{5}$$

$$\boldsymbol{p_\theta(y)} = \boldsymbol{\mathcal{N}(0,I)} \tag{6}$$

$$\boldsymbol{p_\theta(z_t|z_{t-1})} = \boldsymbol{\mathcal{N}(T_\theta^\mu(z_{t-1}), T_\theta^{\sigma^2}(z_{t-1}))} \tag{7}$$

where, f_θ is a feed-forward neural net emission function that acts as a bridge between the latent space and the observation space and T_θ is the gated transition function that is modeled using a feed-forward neural net [7]. The model is reminiscent of a linear dynamical system (LDS), the difference being that the emission and transitions are modeled using non-linear functions f and T respectively.

4.2. FDMM Inference

The goal of inference is to find the *posterior* distribution:

$$\boldsymbol{p(y,z|x)} = \frac{\boldsymbol{p(x,y,z)}}{\boldsymbol{p(x)}} \tag{8}$$

Name	Posterior	Inf. Network
FDMM_i	$q(z_t\|z_{t-1}, x_{t:T})$	RNN & Comb Fxn
FDMM_ii	$q(z_t\|z_{t-1}, x_{1:T})$	BRNN & Comb Fxn
FDMM_iii	$q(z_t\|z_{t-1}, x_{1:t})$	RNN & Comb Fxn

Table 1. Different inference networks. Comb Fxn refers to Combiner Function. See Fig 2 for details

where, $\mathbf{x}$ is the observed variable and $\mathbf{y}, \mathbf{z}$ are the hidden variables. To infer the true posterior distribution is intractable due to the normalization term $p(\mathbf{x})$, hence we turn to variational inference (Section 3). We introduce an approximate posterior $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x})$, which can be written as:

$$q_\phi(\mathbf{z}, \mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} q_\phi(y^n|x^n_{1:T}) \prod_{t=1}^{T} q_\phi(z^n_t|z^n_{t-1}, x^n_{t:T}) \tag{9}$$

where all the posteriors distributions over $\mathbf{y}, \mathbf{z}$ are multivariate diagonal Gaussian distributions. The above factorization can be read off from the inference model structure in Fig 1. Below we give explanation of how individual terms in the right hand side (RHS) of equation 9 are modeled.

The probability density $q_\phi(y^n|x^n_{1:T})$ is given as:

$$q_\phi(y|x_{1:T}) = \mathcal{N}(g^\mu_\phi(x_{1:T}), g^{\sigma^2}_\phi(x_{1:T})) \tag{10}$$

where, g_ϕ is a bidirectional-LSTM (BLSTM) that gives estimates of the parameters of the diagonal Gaussian posterior.

The second term on the RHS $q_\phi(z_t|z_{t-1}, x_{t:T})$ is modeled using two functions, an RNN that takes as input $x_{t:T}$ and provides backward messages at each time step to estimate z_t and a combiner function that combines the backward messages coming from the RNN and the previous estimate $\hat{z}_{t-1}$.

A more pedagogical explanation can be seen in Figure 2. In the figure, given an acoustic segment of length 5, we estimate latent variables z and y as follows. The parameters of the posterior over y is given by a BLSTM encoder followed by two linear transformation layers to give mean and log standard deviation of $q(y|x)$. To get the parameters of the distribution over, say z_2, we combine the backward messages coming from $t = 3, 4, 5$ with the forward message summarized in z_1. Hence, $z_2 \sim q(z_2|z_1, x_{3:5})$. The backward messages are provided by the RNN. This is reminiscent of the *forward-backward* algorithm used to train tradition hidden Markov models and linear dynamical systems. We use the exact same configurations for the RNN backward net and the combiner function as given in [7].

In the following section, we systematically test three forms of the posterior over z_t using different inference networks as given in Table 1.

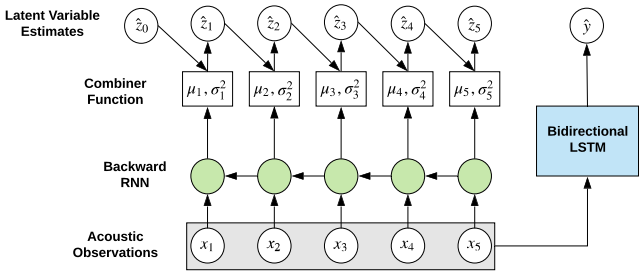

Fig. 2. A pedagogical explanation of the inference process

5. EXPERIMENTS

We use two datasets to conduct the experiments: (a) **TIMIT** [12] which contains 5.4 hours (6,300 utterances), with 10 sentences per speaker, of 16kHz broadband recordings of read speech. 70% of the speakers are male and 30% female. (b) **MGB3** [13] which is a standard dataset used for Arabic dialect identification and consists of 70 hours of 16kHz speech recordings. The data is partitioned into five common dialects; Modern Standard Arabic, Gulf, Levantine, North African and Egyptian.

All speech data is represented as a sequence of 80 dimensional Mel-scale filter bank (FBank) features computed every 10ms. We use the *librosa* [14] toolkit for feature extraction. An observed sample x in our generative model is a 200ms acoustic segment that implies $T = 20$ in the generative model given in Fig 1. i.e. $x \in \mathbb{R}^{20\times 80}$ and latent variables $y, z \in \mathbb{R}^{32}$.

Encoder settings used for experiments are as follows: (a) The encoder for latent variable y is a 2-layered LSTM with 256 hidden units. The output of the encoder is passed through a Gaussian linear layer that outputs the mean and variance estimates for the posterior distribution of y. (b) The encoder for z is a either an RNN or BRNN with hidden units of size 256 in both cases. (c) The combiner function combines the forward, z_{t-1}, and backward, h^{rnn}, messages to give estimates of the conditional posterior distribution of z_t. It first projects the forward message z_{t-1} in the same space as the backward message h^{rnn}_t, then takes an average. The average is then transformed by a linear function to give mean and log standard deviation of $q(z_t|\bullet)$. (d) The transition function T from z_{t-1} to z_t is a gated transition function. Due to lack of space we refer the reader to look at Section 5 of paper [7] for Gated Transition Function. We use the same setup in this work.

Speaker Verification: To evaluate the disentanglement between learned static latent representation y and dynamic z, we perform a speaker verification task using the two latent representations on the TIMIT test data, which contains 24 unique speakers. The data is split into a train, dev, and test and is extracted using the *kaldi* speech recognition toolkit [15]. We use the same exact setup for speaker verification as

Model	Feature	Dim	EER(%)
Factor Analysis	i-vector	200	9.8
FHVAE ($\alpha = 0$)	$\boldsymbol{\mu_2}$	16	5.0
FDMM_i	$\boldsymbol{\mu_y}$	32	6.3
FDMM_i	$\boldsymbol{\mu_z}$	32	18.1
FDMM_ii	$\boldsymbol{\mu_y}$	32	7.0
FDMM_ii	$\boldsymbol{\mu_z}$	32	17.9
FDMM_iii	$\boldsymbol{\mu_y}$	32	5.8
FDMM_iii	$\boldsymbol{\mu_z}$	32	20.0

Table 2. Speaker Verification performance on the TIMIT test data using different feature representations.

used in [5, 8]. Below we give details about feature extraction from our model. These features are used for speaker verification.

For a given speech sequence $\{\boldsymbol{x^n_{1:T}}\}_{\boldsymbol{1:N}}$, where T is the number of frames in the n^{th} acoustic segment for the sequence containing N such segments, we construct two feature representations as follows [8]:

$$\boldsymbol{\mu_y} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{\mu^n_y}, \;\; \boldsymbol{\mu^n_y = E_{q(y^n|x^n_{1:T})}(y^n)} \quad (11)$$

$$\boldsymbol{\mu_z} = \frac{\boldsymbol{1}}{\boldsymbol{TN}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{\mu^n_{z_t}}, \;\; \boldsymbol{\mu^n_{z_t} = E_{q(z^n_t|\bullet)}(z^n_t)} \quad (12)$$

In Table 2 we measure the speaker verification performance in terms of *Equal Error Rate* (EER), a common metric used for this task. Speaker verification proceeds by comparing the feature representation of the test utterance, $\boldsymbol{\mu^{test}}$ with the target utterance $\boldsymbol{\mu^{target}}$ from the claimed identity using *cosine similarity* between the two representations. If the similarity is greater than some threshold ϵ, then the identity is confirmed. Varying ϵ will give us different false acceptance and rejection rates. EER is the point at which false acceptance and rejection rates are equal. For seminal work in speaker verification readers are referred to [16].

Domain Invariant ASR: An open problem in the automatic speech recognition (ASR) community is domain adaptation. Domain adaptation is the problem of adapting a model trained on a data rich domain such as broadcast news to a domain where limited amount of labeled data is available such as for dialects of a language. Feature space domain adaptation refers to the process of extracting domain invariant features from the input signal to then train a model that can be transferred to other domains.

In Table 3 we evaluate whether the latent representations learned by our model contain domain invariant information or not. To that end, we build a phone classification model using TIMIT data. The phone classification model is a 3 layer LSTM [17, 18] with a hidden state size of 1024. We train the model with two input feature types: (a) 80 dimensional Mel-scale FBank features and (b) 32 dimensional $\boldsymbol{z}$ features from the FDMM. To test domain independence we train the phone classification model on utterances spoken by male speakers and test on female speakers; a scenario we consider to be two different "domains". This setup is same as in [5].

		Phone Err. Rate	
Train Data	Feature	Male	Female
All	FBank	25.2	22.0
Male	FBank	27.1	35.8
Male	$\boldsymbol{z}$	27.4	30.1

Table 3. Phone Classification performance using raw FBank features and the latent $\boldsymbol{z}$ features from FDMM

Model	Feature	Dim	ACC(%)
Factor Analysis	i-vector	200	57.4
FHVAE ($\alpha = 10$)	$\boldsymbol{\mu_1}$	32	68.0
FHVAE ($\alpha = 10$)	$\boldsymbol{\mu_2}$	32	54.5
FDMM_iii	$\boldsymbol{\mu_z}$	32	65.2
FDMM_iii	$\boldsymbol{\mu_y}$	32	52.9

Table 4. Dialect ID performance on Arabic dialect id task to test latent space disentanglement

Dialect Identification: To showcase the generality of our model we train an FDMM on the spoken Arabic dialect identification dataset, MGB3. We extract latent features $\boldsymbol{\mu_y}$ and $\boldsymbol{\mu_z}$ in the same way as mentioned in equations 11, 12 and use them to perform five class dialect identification. More details about the task can be found in [19]. Experimental results are shown in Table 4. We use a convolutional neural network based dialect identification system with exactly the same structure as given in [20]. In [21], the authors report dialect identification results using the latent representations learned by an FHVAE and a baseline i-vector system on MGB3. We add these results here for comparison.

In **summary**, speaker verification performance using the dynamic latent variable $\boldsymbol{z}$ is much worse than using $\boldsymbol{y}$ while for domain invariant ASR and dialect identification it is much better. This shows that our model successfully learns to encode content and style/domain information in two different latent variables.

6. CONCLUSION

In this work, we propose a factorial deep Markov model that successfully learns disentangled latent representations from speech. We test our model on three tasks and show that the performance is comparable with an FHVAE model that uses a hierarchical prior to exploit muti-scale information in speech. In the future we hope to investigate more complex models with multiple Markov chains with different time resolutions.

7. REFERENCES

[1] James Glass, "Towards unsupervised speech processing," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 1–4.

[2] Chia-ying Lee and James Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[3] Herman Kamper, "Unsupervised neural and bayesian models for zero-resource speech processing," *arXiv preprint arXiv:1701.00851*, 2017.

[4] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in neural information processing systems*, 2017, pp. 1878–1889.

[6] Wei-Ning Hsu and James Glass, "Scalable Factorized Hierarchical Variational Autoencoder Training," *arXiv preprint arXiv:1804.03201*, 2018.

[7] Rahul G Krishnan, Uri Shalit, and David Sontag, "Structured inference networks for nonlinear state space models.," in *AAAI*, 2017, pp. 2101–2109.

[8] Yingzhen Li and Stephan Mandt, "A Deep Generative Model for Disentangled Representations of Sequential Data," *arXiv preprint arXiv:1803.02991*, 2018.

[9] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[10] Iain Murray, *Advances in Markov chain Monte Carlo methods*, PhD thesis, Gatsby computational neuroscience unit, University College London, 2007.

[11] Michael Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," *arXiv preprint arXiv:1701.02434*, 2017.

[12] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

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

[14] Brian McFee, Matt McVicar, Stefan Balke, Carl Thom, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, WZY, Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis "Fjord" Hawthorne, CJ Carr, Joo Felipe Santos, JackieWu, Erik, and Adrian Holovaty, "librosa/librosa: 0.6.2," Aug. 2018.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[16] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Tenth Annual conference of the international speech communication association*, 2009.

[17] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[18] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[19] Sameer Khurana, Maryam Najafian, Ahmed Ali, Tuka Al Hanai, and Yonatan Belinkov, "Qmdis: Qcri-mit advanced dialect identification system," in *Interspeech*, 2017, pp. 2591–2595.

[20] Suwon Shon, Ahmed Ali, and James Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv preprint arXiv:1803.04567*, 2018.

[21] Suwon Shon, Wei-Ning Hsu, and James Glass, "Unsupervised representation learning of speech for dialect identification," *arXiv preprint arXiv:1809.04458*, 2018.