INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Multi-Accent Acoustic Model using Mixture of Experts for Speech Recognition

Abhinav Jain, Vishwanath P. Singh, Shakti P. Rath

Samsung R&D Institute, India - Bangalore
{jain.abhinav,vp.singh,shakti.rath}@samsung.com

Abstract

A major challenge in Automatic Speech Recognition(ASR) systems is to handle speech from a diverse set of accents. A model trained using a single accent performs rather poorly when confronted with different accents. One of the solutions is a multi-condition model trained on all the accents. However the performance improvement in this approach might be rather limited. Otherwise, accent-specific models might be trained but they become impractical as number of accents increases. In this paper, we propose a novel acoustic model architecture based on Mixture of Experts (MoE) which works well on multiple accents without having the overhead of training separate models for separate accents. The work is based on our earlier work, termed as MixNet, where we showed performance improvement by separation of phonetic class distributions in the feature space. In this paper, we propose an architecture that helps to compensate phonetic and accent variabilities which helps in even better discrimination among the classes. These variabilities are learned in a joint frame-work, and produce consistent improvements over all the individual accents, amounting to an overall 18% relative improvement in accuracy compared to baseline trained in multi-condition style.

Index Terms: multi-accent acoustic model, mixture of experts, deep learning, automatic speech recognition

1. Introduction

With the advancing research in Automatic Speech Recognition and Deep Neural Network (DNN) based Acoustic Models replacing the conventional GMM-HMM models, we have seen significant improvements in ASR performance. This has lead to the application of speech recognition in a plethora of domains, and due to the need for a more natural human computer interaction, these systems have seeped into our daily lives. The immediate consequence is that the ASR technology has to face more challenging problems to support applications in real world scenario, such as speech recognition in far-field and noisy conditions. One of the most important demand in ASR in recent years is to provide support for multiple accents in a unifying framework while improving ASR experience for the users. In literature this is known as multi-accent acoustic modeling.

Dialects[1] are defined as variations within a language that differ in geographical regions and social groups, which can be distinguished by traits of phonology, grammar, and vocabulary [1]. Accents are known to be one of the primary source of speech variability [2]. Due to demographical and social differences, the same language spoken in different parts of the globe leads to significant variations in the way the words are enunciated which poses a serious challenge to current ASR systems. When acoustic model trained over one accent is tested across others, the performance degrades severely due to mismatch in dialects and speaking styles than those observed during system training. Training a separate model for each accent is cumbersome from a commercialization point of view as it increases maintenance cost drastically as number of target accents increase, making seamless support for all accents intractable. This demands for having a unifying frame-work for multi-accent acoustic modeling, the objective of which is to develop a common acoustic model to reduce the word error rates for all target accents simultaneously taking this variability into account. In the simplest form, this is typically achieved by training a single model on all the accents in a multi-condition style simply by pooling data from all target accents. Though promising, the improvement observed in this approach is limited.

In ASR, the essential goal, given a speech signal, is to identify the best possible phone sequence. This task is complicated by the fact that, even in the single accent scenerio, different acoustic classes (such as phonemes) may be strongly overlapped in the acoustic space. Such variations are inherent in natural speech and are difficult to learn, as a result the acoustic model may fail to recognize the correct class especially in the strongly overlapped regions, leading to high inter-class confusion and word recognition errors. On top of that, in the multi-accent scenario, a single phoneme can be enunciated differently by users speaking in different accents. As a result, the overlap in the acoustic space may become more severe due to complex interactions between accent and phonetic variabilities. The objective of the paper is to present an architecture for acoustic modeling that accounts for these two types of acoustic variability in a joint framework.

In our earlier work [3], we developed a novel architecture, named MixNet, for acoustic modeling (in single accent framework) which was based on Mixture of Experts (MoE). MoEs are essentially region-dependent processing of the features using an ensemble of experts, which could be either classifiers or regressors [4, 5, 6]. Different experts are specialized to operate on specific regions in the input space. Outputs of the experts are linearly combined using data dependent weights generated by an additional auxiliary classifier. The role of this classifier is to "select" (soft or hard) the best expert that is akin to the location of the feature vector in the input space [3]. On a large vocabulary ASR task, it was shown that MixNet helps to reduce the overlap in the distribution of different phonetic classes which resulted in a large improvement in the ASR accuracy.

In this paper we propose to extend the previous work by combining two MixNet layers sequentially, in which the first layer accounts for phonetic variability and the other accounts for accent variability. It is found that MixNet used alone for phonetic variability gives 6% improvement relative, while MixNet used only for accent compensation gives 7.5% improvement separately. We show that our proposed architecture which jointly learns phonetic and accent variabilities helps in

[1] In this paper we use the terms dialect and accent interchangeably.

http://dx.doi.org/10.21437/Interspeech.2019-1667

better discrimination among the classes increasing the overall improvement to 18% relative.

The organization of the rest of the paper is as follows. In Section 2, we present a survey on multi-accent speech recognition. Sections 3 and 4 will look into *MixNet* and our proposed approach. Description of data used and experimental setup are discussed in Section 5 and Section 6, respectively. Section 7 presents summary and scope of future work.

2. Related Work

Accented speech recognition has been an extensively researched in earlier years. One of the most obvious and earliest techniques was to use an augmented dictionary, i.e. adding accent specific pronunciations, which reduced cross-accent recognition error rates [7]. Apart from English language, multiple accents in languages like Chinese and Afrikaans have also been studied in prior work [8, 9, 10, 11].

Initial approaches for accented speeh recognition on adapting acoustic and pronunciation models were based on GMM-HMM based models [8, 9, 10, 12]. With time, DNN based Acoustic Models [13] became a standard and DNN-based adaptation approaches included having accent-specific output layers and shared hidden layers [11, 14]. The use of model interpolation has also been utilized to learn accent dependent models where the interpolation weights were learnt using data [15]. In a more recent work, connectionist temporal classification (CTC) loss function was proposed for multi-accented speech [16] where hierarchical grapheme-based models that jointly predicts both graphemes and phones performed the best. Recently, leveraging native language information for accented speech recognition [17] have proved to be effective where the authors use native language data in a multitask framework. Multitask learning has also been used to jointly learn accent classifier and acoustic model. Yang et al. [1] used an Acoustic Model for American and British Accented Speech and model selection using accent classifier whereas Jain et al. [18] proposed Accent Embeddings and a generalized multi-accent Acoustic Model which is shown to work on both seen and unseen accents.

3. Review of MixNet for Phonetic Variability

In our previous work [3], we developed *MixNet*, which uses mixture of experts (MoE) to "segregate" phonetic classes in the feature space (in a single accent scenario). Specifically, in one of the four configurations in the paper (i.e., MixNet-I), the input features are transformed by an MoE layer in which region-dependent (experts) affine transforms are applied to move the features in different directions in the acoustic space. Experimentally it was verified that it leads to better separation of the phonetic-classes in the feature space and resulted in large improvement in ASR accuracy.

The block diagram of MixNet-I is shown in Figure 1. The experts, denoted by $\mathbf{A}_i$ matrices in the Figure, are learned on top of input features belonging to different acoustic regions. The regions may be pre-defined to be broad phonetic classes. An auxiliary classifier is trained to classify input features into these classes. The outputs from the experts are then combined linearly using posterior probabilities generated by a classifier as

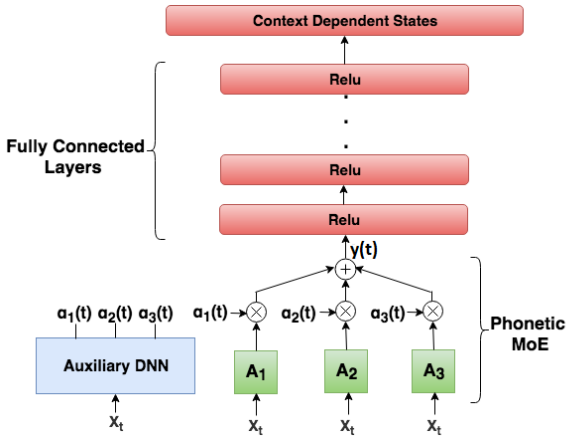

Figure 1: *Schematic diagram of MixNet: Mixture of Experts for phonetic variability*

weights. Mathematically,

$$\mathbf{y}(t) = \sum_{i=1}^{C_p} \alpha_i(t)\mathbf{A}(i)\mathbf{x}(t) \quad (1)$$

where C_p is the number of the broad phonetic classes, $\mathbf{x}_t$ and $\mathbf{y}_t$ are the input and output features of the MoE network. $\alpha_i(t)$ is the posterior probability (or the gating signal) of class i pertaining to frame at time t, which is generated by the classifier.

4. Proposed Architecture for Joint Phonetic and Accent Variability

As discussed in the Introduction, in a multi-accent scenario, the speech is affected by highly non-linear interaction between phonetic and accent variability that may lead to strongly overlapped phonetic classes. In this paper, we propose an extension of MixNet where we aim not only to account for phonetic variability but also accent variability in a joint frame-work. The developments are done in steps, where we first consider MixNet only for phonetic variability, followed by MixNet only for accent variability, followed by MixNet for both.

We start with a Time-Delay Neural Network (TDNN) [19] as the baseline Acoustic Model. The use of TDNN is motivated by the fact that it performs at par with recurrent neural network based acoustic models [19, 20] while still being able to be trained in parallel due to the absence of recurrent connections. They are successful in learning long-term dependencies in the input signal using only short-term acoustic features. This is a single model trained on data containing multiple accents. We extend this TDNN-HMM model to incorporate MixNet layers. With this aim two MixNet layers are applied (Figure 2), the first one designed to compensate phonetic variability and the second one for accent variability. It is argued that the proposed architecture helps to compensate phonetic and accent variabilities which learns better discrimianatory features, hence improving the performance for the various accents.

4.1. MixNet for Phonetic Variability: MixNet-P

As proposed in [3] and described in Section 3, we apply a MixNet layer just before the first TDNN layer for phonetic variability. The "experts" in this MoE layer are affine transforms operating on 3 phonetic classes, namely *voiced*, *unvoiced* and

silence. The experts are controlled by "gating signals" generated by an auxiliary TDNN Phone Classifier **(PhoneAuxNet)**. This auxiliary TDNN is trained to classify acoustic feature vectors into these 3 phonetic classes.

4.2. MixNet for Accent Variability: MixNet-A

In the second extension, a MixNet layer is trained to transform the input features in the accent space. The architecture is similar to MixNet-P, with the difference that the auxiliary classifier is now trained to classify frames into accents. The "experts" are thus specific to regions in the accent space, instead of phonetic classes. This MixNet layer is inserted just before the first TDNN layer. Similar to MixNet-P, these experts are controlled by the outputs of the auxiliary TDNN Accent Classifier **(AccentAuxNet)**. Mathematically,

$$\mathbf{z}(t) = \sum_{i=1}^{C_a} \beta_i(t) \mathbf{B}_i \mathbf{x}(t) \tag{2}$$

where C_a is the number of accent classes, in our case 9 (8 accents and 1 additional class for silence). $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are the input and output of the MixNet layer respectively and $\beta_i(t)$ is the posterior probability of class i for frame at time t. $\mathbf{z}(t)$ is fed to the following TDNN layers in the acoustic model.

4.3. MixNet for Phonetic & Accent Variability: MixNet-PA

The major contribution of this paper is a novel architecture which uses both phonetic and accent transformation as described in Sections 4.1 and 4.2 in joint frame-work. The architecture of the network is shown in Figure 2.

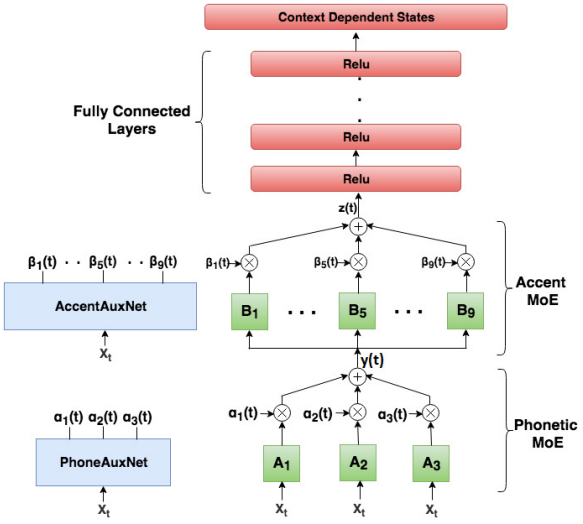

Figure 2: *Schematic diagram of MixNet-PA for joint Phonetic and Accent Variability*

The model uses two MixNet layers sequentially. The experts in the first layer learn transformation of the input acoustic feature that will increase separation of the phonetic classes. These experts are represented by $\mathbf{A_i}$ matrices. The operation is given by Eq. 1. This is followed by MixNet-A which acts upon $\mathbf{y}(t)$ and learns transformations specific to accents in the accent space (Eq. 2, except $\mathbf{x}(t)$ is replaced by $\mathbf{y}(t)$).

5. Data Description

We use a multi-accent set of SpeechOcean[21] data for all our experiments. The data set included a total of 8 accents of English namely, Australian (AUS), Canadian (CAN), British (GB), Indian (IN), Korean (KR), Chinese (CHN), United States (US) and Spanish (SPN). The training set ("TRAIN"), the validation set ("VAL") and the test set ("TEST") are a mixture of these 8 accents. The detailed statistics are presented in Table 1.

Table 1: *Statistics of the data used in all the experiments.*

Dataset	Hrs. of Speech	# Sentences
TRAIN	736	490792
DEV	1.2	800
TEST	6	4000

TRAIN is used for training all the networks. *DEV* is used for hyperparameter tuning and all the results in the paper are reported on *TEST*. *TEST* and *DEV* are equal in accent representation and the composition of *TRAIN* in given in Table 2.

Table 2: *Accent-wise break-down of the TRAIN set.*

Accent	Percent	Hrs. of Speech	# Sentences
AUS	14.3	105	70000
KR	7	50	33746
CAN	7.2	56	37266
SPN	14.3	105	70000
CHN	14.3	105	70000
US	14.3	105	69780
GB	14.3	105	70000
IN	14.3	105	70000

6. Experimental Analysis

6.1. Baseline

All experiments are conducted using the Kaldi toolkit [22]. The first baseline system is a feed-forward TDNN network with subsampling at intermediate layers. The first layer performs affine transform on spliced frames (frames forming a window of size $t-2$ to $t+2$). Following the input layer, the network consists of 7 layers of 1024 nodes with ReLU activation function spliced with offsets {0}, {-1,2}, {-3,3}, {-5,3}, {-7,2}, {-9,4}, {0} respectively with cross entropy loss across senones. We use 41-dim fbank features as input. For decoding, a tri-gram language model was trained using the training transcripts. This model is trained in a multi-condition style where speech from all accents are pooled. We refer to this model as *Baseline* in the experiments.

For comparison, the results are shown for a second baseline network proposed in [18], which is a multitask network. It was shown that this model improves accented speech recognition accuracy using accent dependent features learned end-to-end. This network is referred to as *Multitask Baseline* in the experiments.

6.2. Auxiliary Classifier Networks

The details of the two classifiers used in our experiments namely *PhoneAuxNet* and *AccentAuxNet* are given below.

PhoneAuxNet: The classifier used in conjunction with

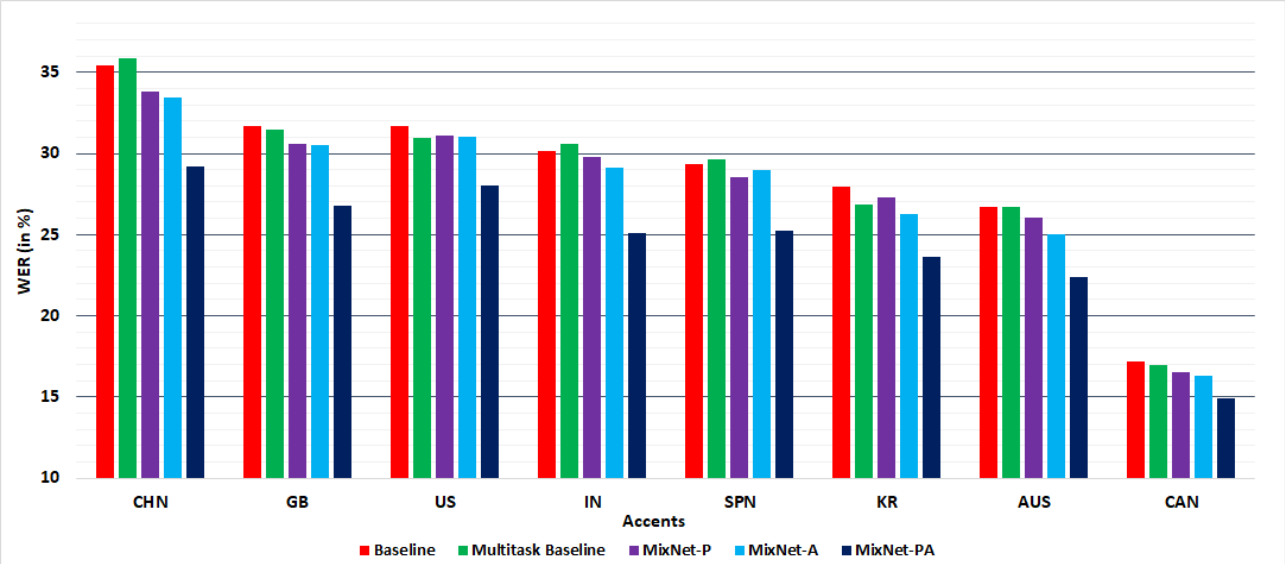

Figure 3: *Accent-wise WERs on TEST set of all the proposed networks. MixNet-PA gives substantial improvement on every accent.*

MixNet-P layer of the network is a 4-layer TDNN phone classifier consisting of 512 nodes with ReLU activation function spliced with offsets {-1,2}, {-3,3}, {-3,3}, {0} respectively. As discussed, the classes are 3 broad phonetic classes. The targets are obtained by forced alignment of the utterances and mapping each phone to its corresponding class. We experimented with higher number of phonetic classes, but the performance was optimum for 3-classes.

AccentAuxNet: The classifier used with MixNet-A layer consisted of 8 accent classes plus one additional class for non-speech phones (to account for non-speech frames). The network comprises of 5 layers of TDNN containing 512 nodes each with ReLU activation function. The first layer takes the current frame and 2 frames each, as left and right context, followed by layers with offsets {-1,2}, {-3,3}, {-7,2} , {0} respectively. The targets are (frame level) accents for each utterance. The frame-level accuracy of the two classifiers are shown in Table 3.

Table 3: *Frame accuracy of auxiliary networks on the VAL set*

Model	Accuracy (%)
PhoneAuxNet (3 class)	87.0
AccentAuxNet (9 class)	92.0

6.3. Improvement using MixNet-P and MixNet-A

Table 4 shows the recognition performance of the network architectures proposed in Sections 4.1 and 4.2 compared to our two baseline systems. It can be seen that both MixNet-P and MixNet-A give a relative improvement of 6% and 7.5% respectively over the first baseline whereas they still perform poorly as compared to our *Multitask Baseline*.

6.4. Improvement using MixNet-PA

We finally explore the performance of the proposed MixNet-PA which combines both MixNet-P and MixNet-A sequentially. Table 5 shows the WER in comparison with the two baselines.

Table 4: *Performance of MixNet-P & MixNet-A*

Model	Parameters	WER (%)
Baseline	20.5M	30.2
Multitask Baseline	29.2M	27.2
MixNet-P	22.5M	28.3
MixNet-A	23M	27.9

It is observed that MixNet-PA outperforms both our Baseline networks by considerable margins. Specifically, it is better than the first Baseline by 5.5% absolute which corresponds to 18.2% relative. Moreover, the improvement is consistent over all the accents as shown in the plot (Figure 3). It is also noted that MixNet-PA contains much less number parameter than Multitask Baseline, yet the performance is significantly better.

Table 5: *WERs (in %) of MixNet-PA compared to the baselines.*

Model	Parameters	WER (%)
Baseline	20.5M	30.2
Multitask Baseline	29.2M	27.2
MixNet-PA	26M	**24.7**

7. Conclusions

In this work, we explore the use of Mixture of Experts based acoustic model for multi-accent speech recognition where we create a unified acoustic model for multiple accents. This network learns to segregate accent-specific and phone-specific speech variabilities in a joint frame-work, which gives far superior performance compared to a multi-accent baseline system, obtaining upto 18% relative WER reduction on the test set. For future work, we will investigate different architectures and analyze the segregation of the input features. We will also look into tying similar accents into one expert.

8. References

[1] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[2] C. Huang, T. Chen, S. Z. Li, E. Chang, and J.-L. Zhou, "Analysis of speaker variability," in *Proceedings of Eurospeech*, 2001.

[3] V. P. Singh and S. P. Rath, "Mixnet: A mixture of expert based deep neural network for improved asr," *IEEE Signal Process. Lett.*, Submitted for Publication.

[4] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.

[5] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton *et al.*, "Adaptive mixtures of local experts." *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[6] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[7] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proceedings of ICSLP*, vol. 4. IEEE, 1996, pp. 2324–2327.

[8] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Proceedings of European Conference on Speech Communication and Technology*, 2005.

[9] Y. Liu and P. Fung, "Multi-accent chinese speech recognition," in *Proceedings of ICSLP*, 2006.

[10] H. Kamper and T. Niesler, "Multi-accent speech recognition of afrikaans, black and white varieties of south african english," in *Proceedings of Interspeech*, 2011.

[11] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in *Proceedings of Interspeech*, 2015.

[12] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented english data," in *Proceedings of Interspeech*, 2010.

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[14] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Proceedings of Interspeech*, 2014.

[15] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Speech recognition of multiple accented english data using acoustic model interpolation," in *Proceedings of EUSIPCO*. IEEE, 2014, pp. 1781–1785.

[16] K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *Proceedings of ICASSP*. IEEE, 2017, pp. 4815–4819.

[17] S. Ghorbani and J. H. Hansen, "Leveraging native language information for improved accented speech recognition," *Proc. Interspeech 2018*, pp. 2449–2453, 2018.

[18] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Proc. Interspeech 2018*, 2018, pp. 2454–2458. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1864

[19] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[20] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Proceedings of Interspeech*, 2015.

[21] SpeechOcean, "Kingline Data Center." [Online]. Available: https://kingline.speechocean.com

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.