Interspeech 2018
2-6 September 2018, Hyderabad

Dysarthric Speech Recognition using Time-delay Neural Network based Denoising Autoencoder

Chitralekha Bhat, Biswajit Das, Bhavik Vachhani, Sunil Kumar Kopparapu

TCS Research and Innovation, Mumbai, India

{bhat.chitralekha, b.das, bhavik.vachhani, sunilkumar.kopparapu}@tcs.com

Abstract

Dysarthria is a manisfestation of the disruption in the neuromuscular physiology resulting in uneven, slow, slurred, harsh or quiet speech. Dysarthric speech poses serious challenges to automatic speech recognition, considering this speech is difficult to decipher for both humans and machines. The objective of this work is to enhance dysarthric speech features to match that of healthy control speech. We use a Time-Delay Neural Network based Denoising Autoencoder (TDNN-DAE) to enhance the dysarthric speech features. The dysarthric speech thus enhanced is recognized using a DNN-HMM based Automatic Speech Recognition (ASR) engine. This methodology was evaluated for speaker-independent (SI) and speaker-adapted (SA) systems. Absolute improvements of 13% and 3% was observed in the ASR performance for SI and SA systems respectively as compared with unenhanced dysarthric speech recognition.

Index Terms: Time-Delay Neural Network, Deep denoising autoencoders, Dysarthric Speech, Speech Enhancement

1. Introduction

Speech production process comprises acoustic and linguistic events that occur through the coordination of muscle groups and neurological programming of muscle activities, to ensure fluent and accurate articulation. Acquired or developmental dysarthria, results from the impairment of the motor execution function and affects the speech itelligibility of a person. Voice input-based interactions with smart devices perform poorly for dysarthric speech. Research into automatic recognition of dysarthric speech has garnered much interest due to the rising popularity and possibility of voice inputs, especially since speech-based interaction is easier for persons with neuro-motor disorders as compared to keypad inputs [1].

Several techniques are employed to improve ASR performance for dysarthric speech: acoustic space enhancement, feature engineering, Deep Neural Networks (DNN), speaker adaptation, lexical model adaptation- individually or as a combination thereof. Formant re-synthesis preceded by modifications of formant trajectories and energy, for dysarthric speech vowels showed significant improvement in perceptual evaluation of intelligibility of CVC utterances [2]. Acoustic space modification carried out through temporal and frequency morphing improved automatic dysarthric speech recognition as well as subjective evaluation in [3]. It can be seen that temporal adaptation based on dysarthria severity level improved the ASR performamce for dysarthric speech recognition at each severity level [4]. A Convolutive Bottleneck Network (CBN) was used for dysarthric speech feature extraction wherein the pooling operations of the CBN resulted in features that were more robust towards the small local fluctuations in dysarthric speech and outperformed the traditional MFCC feature based recognition [5]. A comparative study of several types of ASR systems including maximum likelihood and maximum a posteriori (MAP) adaptation showed a significant improvement in dysarthric speech recognition when speaker adaptation using MAP adaptation was applied [6]. Word error rate for dysarthric speech was reduced using voice parameters such as jitter and shimmer along with multi-taper Mel-frequency Cepstral Coefficients (MFCC) followed by speaker adaptation [7], and using Elman back-propagation network (EBN) which is a recurrent, self supervised neural network along with glottal features and MFCC in [8]. A multi-stage deep neural network (DNN) training scheme is used to better model dysarthric speech, wherein only a small amount of in-domain training data showed considerable improvement in the recognition of dysarthric speech [9]. In [10], authors propose a DNN based interpretable model for objective assessment of dysarthric speech that provides users with an estimate of severity as well as a set of explanatory features. Speaker selection and speaker adaptation techniques have been employed to improve ASR performance for dysarthric speech in [11, 12]. ASR configurations have been designed and optimized using dysarthria severity level cues in [13, 14, 15].

It has been observed that the subjective perception-based intelligibility performance for noisy and dysarthric speech is correlated, indicating that there exists similarity in the information processing of these two types of speech [16]. Extrapolating this to the objective assessment domain, we hypothesize that techniques used for noisy speech may support dysarthric speech processing as well. In this paper we explore the possibility of using a Time-Delay Neural Network Denoising Autoencoder (DAE) for dysarthric speech feature enhancement. DAEs have been used to enhance speech features especially in noisy conditions [17, 18, 19]. The objective is for the network to learn a mapping between dysarthric speech features and the healthy control speech features. This network is then used to enhance the dysarthric speech features that are used in a DNN-HMM based ASR for improved dysarthric speech recognition. ASR performance indicates that the enhanced dysarthric speech features are closer to healthy control speech features rather than dysarthric speech features. Evaluation of our work is carried out on Universal Access Dysarthric Speech corpus [20]. In our earlier work [21], we had used a Deep Autoencoder to enhance dysarthric test speech features, wherein the DAE was trained using only healthy control speech. This is different from our current work in the DAE configuration and the training protocol followed.

The rest of the paper is organized as follows. Section 2 describes the methodology employed to enhance speech features for dysarthric speech recognition, Section 3 discusses the experimental setup, In Section 4 we discuss the results of our experiments we conclude in Section 5.

10.21437/Interspeech.2018-1754

2. Dysarthric Speech Feature Enhancement

The process and techniques used to enhance dysarthric speech features is described in this Section.

2.1. Time-Delay Neural Network

TDNN architecture is capable of representing relationships between events in time using a feature space representation of these events [22]. Computation of the relationship between current and past inputs is made possible by introducing delays to the basic units of a traditional neural network as shown in Figure 1.

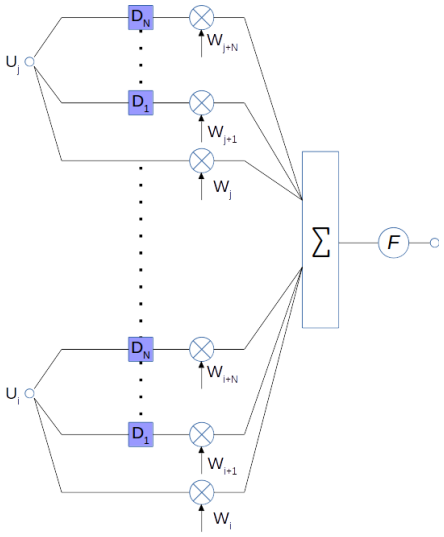

Figure 1: *Time delay neural network unit [22]*

The discovery of the acoustic features and the temporal relationship between them independent of position of time ensures that the dysarthric speech features are not blurred by the inherent small local fluctuations. Shorter temporal contexts are used to learn the initial transforms whereas the hidden activations from longer contexts are used to train the deeper layers. This enables the higher layers to learn longer temporal relationships [23].

Back-propagation learning is used to train TDNN-DAE, wherein the input features are extracted from noisy speech and target features are extracted from the corresponding clean speech.

2.2. Methodology

In traditional DAE training, the number of frames in the input utterance must necessarily be equal to the number of frames in the target utterance. This works well for scenarios wherein noise added clean speech is the input and the corresponding clean speech is the target. In this work, we intend to use dysarthric speech as input and its healthy control counterpart as the target speech, since the objective is for the TDNN-DAE network to learn the mapping between the two. Typically dysarthric speech is slower than healthy control speech and hence of longer duration. One mechanism to match the number of frames is by using varying frame lengths and frame shifts for dysarthric utterance so as to match the number of frames in the corresponding healthy control utterance. However, the difference in the durations between dysarthric utterances and healthy control utterances was too high to achieve a meaningful frame lengths and frame shifts.

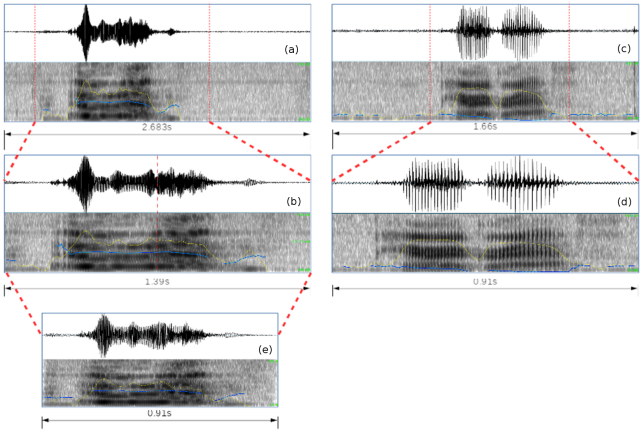

Figure 2: *Data Preparation for TDNN-DAE training for the word 'Paragraph'- (a) Original dysarthric utterance (2.68s) (b) Dysarthric utterance after end point silence removal (1.39s) (c) Original healthy control utterance of duration (1.66s) (d) Healthy Control utterance after end point silence removal (0.91s) (e) Dysarthric utterance after tempo adaptation (0.91s) to match (d)*

Matching of number of frames was done using the following two steps as depicted in Figure 2.

- Majority of silence portion at the beginning and ending of both dysarthric and healthy control utterances was eliminated retaining roughly 200 ms of silence.

- In order to match the durations of the input dysarthric utterance and target healthy control utterance, the dysarthric utterance was temporally adapted using phase vocoder as described in [3]. Tempo adaptation is carried out according to the adaptation parameter α given as $\alpha = \frac{d_H}{d_D}$ where d_D is the duration of the dysarthric utterance and d_H is the duration of healthy control utterance. Tempo adaptation using phase vocoder based on short-time Fourier transform (STFT) ensures that the pitch of the sonorant regions of dysarthric speech is unaffected [24]. Magnitude spectrum and phase of the STFT are either interpolated or decimated based on the adaptation parameter (α), where the magnitude spectrum is directly used from the input magnitude spectrum and phase values are chosen to ensure continuity. This ensures that the pitch of the time-warped sonorant region is intact. For the frequency band at frequency f and frames i and $j > i$ in the modified spectrogram, the phase Θ is predicted as

$$\Theta_j^f = \Theta_i^f + 2\pi f \cdot (i - j) \quad (1)$$

  The modified magnitude and phase spectrum are then converted into a time-domain signal using inverse Fourier transform.

Figure 3 shows the proposed methodology for a TDNN-DAE based dysarthric speech feature enhancement and recognition.

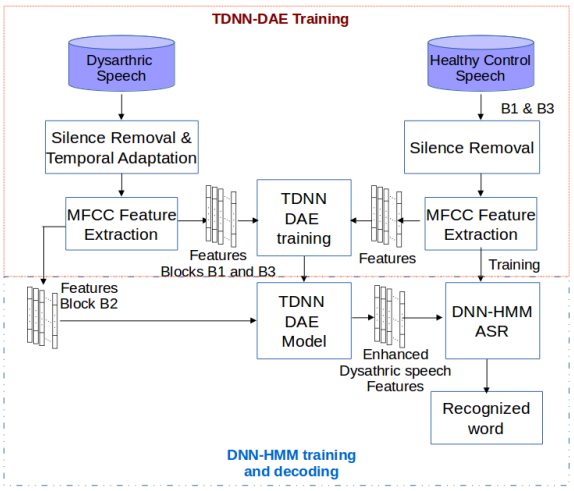

Figure 3: *TDNN-DAE based dysarthric speech feature enhancement and recognition.*

3. Experimental Setup

TDNN-DAE as well as DNN-HMM based ASR were implemented using Kaldi speech recognition toolkit [25].

3.1. Dysarthric Speech Corpus

Data from Universal Access (UA) speech corpus [20] was used for training the TDNN-DAE and DNN-HMM based ASR systems. UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Data was collected in three separate sessions for each speaker and categorized into three blocks B1, B2 and B3. In each block a speaker recorded 455 distinct words and a total of 765 isolated words. The corpus also includes speech intelligibility rating for each dysarthric speaker, as assessed by five naive listeners.

3.2. TDNN-DAE

23 dimensional Mel-frequency cepstral coefficients (MFCC) were used as input features for all the experiments. TDNN-DAE architecture described in [23] was followed. Contexts for the DAE network with 4 hidden layers is organized as (-2,-1,0,1,2) (-1,2) (-3,3) (-7,2) (0) which is asymmetric in nature. Input temporal context for the network is set to [-13,9]. It can be observed that narrow context is selected for initial hidden layers whereas higher contexts for deeper layers. Each hidden layer comprises 1024 ReLU activation nodes. TDNN-DAE was trained using training data described in Section 3.1.

3.2.1. Training data

In this work, we use 19 computer command (CC) words from blocks B1 and B3 of dysarthric speech and of healthy control speech for TDNN-DAE training. Each dysarthric utterance was temporally adapted with each of its corresponding healthy control utterance. For example the dysarthric utterance F05_B1_C12_M2.wav (spoken by speaker F05 recorded as block B1 on channel M2) corresponding to CC word *C12:Sentence*, was temporally adapted to match the duration of each of the healthy control utterance corresponding to the CC word *C12:Sentence*. Thus generating multiple dysarthric utterances from one single dysarthric utterance as shown in Equation given below.

$$D_{u_i j} = f(D_{u_i j}, \forall_{H_{u_i}} H_{u_i}) \quad (2)$$

where $u_i \rightarrow$ CC utterances with $i = 1 \cdots 19$
$D_{u_i j} \rightarrow$ dysarthric utterance where $j = 1 \cdots 3511$
$H_{u_i} \rightarrow$ healthy control CC utterances with $i = 1 \cdots 19$
$f \rightarrow$ temporal adaptation(TA) function [4]

A total of 3511 dysarthric utterances were temporally adapted against their healthy control counterparts generating around 0.6 million temporally adapted dysarthric utterances. The TDNN-DAE was trained using the temporally adapted dysarthric speech utterances as input speech while their corresponding healthy control utterances comprised the target speech.

3.2.2. Testing data

TDNN-DAE trained as above was used to enhance the dysarthric speech features corresponding to 1791 utterances `i.e.` computer command words from block B2. These utterances were first temporally adapted followed by enhancement of the corresponding MFCC features using TDNN-DAE. These enhanced speech features for dysarthric speech were used to evaluate ASR recognition performance.

3.3. DNN-HMM based ASR

Dysarthric speech was recognized using the same configuration of DNN-HMM as in our previous work [21]. A maximum likelihood estimation (MLE) training approach with 100 senones and 8 Gaussian mixtures was adopted. Cepstral mean and variance normalization (CMVN) was followed by dimensionality reduction using Linear Discriminant Analysis (LDA) with a context of 6 frames (3 left and 3 right) to give a feature vector of size 40. The input layer of DNN has 360 (40×9 frames) dimensions. Two hidden layers with 512 nodes in each layer and an output layer of dimension 96 were used. A constrained Language Model (LM), wherein we restrict the recognizer to give one word as output per utterance was used.

Healthy control (HC) and dysarthric (DYS) speech utterances from blocks B1 and B3 of computer command (CC) words were used for training the DNN-HMM based ASR as shown in Table 1. Training configuration S-1 comprises only healthy control (HC) speech. In the second training configuration S-2, we use dysarthric (DYS) speech from blocks B1 and B3 in addition to HC speech. In S-3, ASR was trained using HC speech and dysarthric speech from blocks B1 and B3 that were enhanced using the TDNN-DAE models. Each training configuration was evaluated using dysarthric speech features for computer command words (DYS) from block B2. In Testing configuration 1, the dysarthric speech features were temporally adapted. In our earlier work [4], we show that temporal adaptation of the test dysarthric speech significantly reduced the ASR word error rate (WER). Hence, in this paper we use the WER corresponding to temporally adapted dysarthric speech as baseline. In Testing configuration 2, the temporally adapted dysarthric speech features were enhanced using the TDNN-DAE model and then evaluated. There is no overlap in the training and testing data.

4. Results and Analysis

DNN-HMM ASR recognition is evaluated for speaker adaptation (SA) and speaker independent (SI) scenarios for the train-

Table 1: *ASR Training and testing configurations*

System	**Training configuration** (B1,B3)	**Testing configuration 1** (B2)	**Testing configuration 2** (B2)
S-1	HC	Temporally adapted DYS (MFCC-TA)	Temporally adapted + TDNN-DAE enhanced DYS (MFCC-TA+TDNN-DAE)
S-2	HC + DYS		
S-3	HC + TDNN-DAE enhanced-DYS		

ing and test cofigurations mentioned in Table 1. Word error rates produced for the above scenarios are reported in Table 2. System S-1 does not use any dysarthric speech data for ASR training. An absolute improvement of 13% was observed when the test dysarthric speech data was enhanced using the TDNN-DAE. This indicates that the TDNN-DAE based enhancement of dysarthric speech features results in these features being closely matched to healthy control speech features. Also, the drastic reduction in the ASR performance for S-2 for TDNN-DAE enhanced data, specifically in the SA scenario serves as additional confirmation that the enhanced dysarthric speech features match more closely to healthy control than to dysarthric speech data. Training configuration S-3 comprises healthy control and TDNN-DAE enhanced dysarthric data (B1 and B3). Speaker adaptation based ASR performance is higher by 3% for TDNN-DAE enhanced dysarthric speech (B2) than SA recognition performance for S-2. Both S-2 and S-3 contain the same amount of healthy control and dysarthric speech data in the training process, except that the dysarthric speech used in S-3 is enhanced using TDNN-DAE. ASR performance for the three different training configurations clearly indicates that using TDNN-DAE to enhance dysarthric speech features results in dysarthric speech features matching closely to healthy control speech.

Table 2: *WER for TDNN-DAE*

Training configuration	Testing configuration 1		Testing configuration 2	
	SA	SI	SA	SI
S-1	-	37.86	-	24.73
S-2	21.44	33.67	60.8	29.7
S-3	82.69	72.47	18.54	34.39

An analysis of ASR performance at dysarthria severity levels was done for the two configurations that provide the best recognition, namely S-2-SA using unenhanced dysarthric training and test data and S-3-SA using enhanced dysarthric training and test data. An improvement was seen across all Dysarthria severity levels.

Table 3: *Severity level analysis of WER*

Severity	S-2-SA Testing configuration 1	S-3-SA Testing configuration 2	Absolute Improvement
Very Low	5.71	1.35	4.4
Low	11.39	9.4	1.99
Medium	22.67	19.46	3.2
High	57	52.5	4.5

5. Conclusion

In this paper we explain the process of enhancing dysarthric speech features using a TDNN-DAE. The objective is to enhance the dysarthric speech features to match that of healthy control speech. TDNN-DAE is trained using temporally adapted dysarthric speech as input and healthy control speech as target speech. The training process and the data used for TDNN-DAE need careful consideration to obtain optimal ASR performance. The dysarthric speech thus enhanced is recognized using a DNN-HMM based Automatic Speech Recognition (ASR). Speaker independent and speaker adaptation based ASR configurations were evaluated using both unenhanced and enhanced dysarthric. An absolute improvement of 13% and 3% was observed in ASR performance for SI and SA configurations respectively when enhanced dysarthric speech features were used. ASR performance for each of the training and testing configurations confirm that the dysarthric speech enhanced using TDNN-DAE is matched more closely to healthy speech than to dysarthric speech for the same speaker. An analysis of the two best performing configurations clearly indicate that the ASR performance significantly improves at all severity levels of dysarthria.

6. References

[1] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. NIPS*, 2010, pp. 70–78.

[2] A. B. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2007.05.001

[3] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013, special Issue on SLPAT.

[4] C. Bhat, B. Vachhani, and S. Kopparapu, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*, 2016, pp. 370–377.

[5] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *2014 12th International Conference on Signal Processing (ICSP)*, Oct 2014, pp. 505–509.

[6] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. Interspeech*, 2012, pp. 1776–1779.

[7] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation," in *Proc. Interspeech*, 2016, pp. 228–232.

[8] S. Selva Nidhyananthan, R. Shantha Selva kumari, and V. Shenbagalakshmi, "Assessment of dysarthric speech using elman back propagation network (recurrent network) for speech recognition," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 577–583, Sep 2016.

[9] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multi-stage dnn training for automatic recognition of dysarthric speech," in *Proc. Interspeech*, 2017, pp. 2685–2689.

[10] J. L. Ming Tu, Visar Berisha, "Interpretable objective assessment of dysarthric speech based on deep neural networks," in *Proc. Interspeech*, 2017, pp. 1849–1853.

[11] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 254–259.

[12] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147 – 1162, 2013, special Issue on SLPAT.

[13] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *SLPAT*, 2015, p. 65.

[14] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. Siong, "Severity-based adaptation with limited data for ASR to aid dysarthric speakers," *PLoS One*, 2014.

[15] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models." in *Proc. Interspeech*, 2013, pp. 3622–3626.

[16] S. A. Borrie, M. Baese-Berk, K. Van Engen, and T. Bent, "A relationship between processing speech in noise and dysarthric speech," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4660–4667, 2017.

[17] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *Proc. Interspeech*, 2016, pp. 3743–3747.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. Interspeech*, 2013, pp. 436–440.

[19] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *ICASSP*, May 2014, pp. 1759–1763.

[20] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *Proc. Interspeech*, 2008, pp. 1741–1744.

[21] B. Vachhani, C. Bhat, B. Das, and S. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *Proc. Interspeech*, 2017, pp. 1854–1858.

[22] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *IEEE Trans. on Acoustics, Speech, and Signal Processing*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 393–404.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[24] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, Jun 1976.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.