



Automatic Speaker Verification System for Dysarthria Patients

Shinimol Salim¹, Syed Shahnawazuddin², Waquar Ahmad³

^{1,3}Dept of Electronics and Communication Engineering, NIT Calicut, India

²Dept of Electronics and Communication Engineering, NIT Patna, India

shinimol.p190017ec@nitc.ac.in, s.syed@nitp.ac.in, waquar@nitc.ac.in

Abstract

Dysarthria is one of the most common speech communication disorder associate with a neurological damage that weakens the muscles necessary for speech. In this paper, we present our efforts towards developing an automatic speaker verification (ASV) system based on x -vectors for dysarthric speakers with varying speech intelligibility (low, medium and high). For that purpose, a baseline ASV system was trained on speech data from healthy speakers since there is severe scarcity of data from dysarthric speakers. To improve the performance with respect to dysarthric speakers, data augmentation based on duration modification is proposed in this study. Duration modification with several scaling factors was applied to healthy training speech. An ASV system was then trained on healthy speech augmented with its duration modified versions. It compensates for the substantial disparities in phone duration between normal and dysarthric speakers of varying speech intelligibility. Experiment evaluations presented in this study show that proposed duration-modification-based data augmentation resulted in a relative improvement of 22% over the baseline. Further to that, a relative improvement of 26% was obtained in the case of speakers with high severity level of dysarthria.

Index Terms: automatic speaker verification system, dysarthric speech, data augmentation, duration modification, x -vector

1. Introduction

Speech production is a complex mechanism that requires a coordinated and timely contraction of an expansive number of muscle bunches related to respiration, laryngeal control, and articulation [1]. Neuro-muscular conditions that affect the nerves controlling these muscles can cause debilitation of speech intelligibility and quality as well as the innate speech characteristics of individual speaker. Dysarthria refers to a category of neurological speech disorders in which part of the brain that controls the muscles involved in speech production is either weakened, injured, or damaged [2]. Some of the distinct expressions of dysarthria include slurred speech, slow speech with destitute articulation [3]. It is also linked to issues with both excitation and vocal tract setting, speed of changing the position of articulators, etc. Problems affecting the larynx change the quality of phonation, pitch, and loudness of speech [4]. The speaker may have shallow breathing, difficulty adjusting exhalation with vocalization. The involvement of the soft palate usually leads to the perception of excessive nasal sounds in dysarthric speech. Dysarthria can range in severity from mild to severe. As the condition becomes more and more severe, speech is rendered nearly unintelligible [5].

Voice biometric technologies have grown in prominence over the past few decades. Compared to existing biometric technologies, voice authentication is more adaptable, accurate, and non-intrusive [6]. The specific recognition task that a commer-

cial system deals with is a verification task, rather than an identification task. An automatic speaker verification (ASV) system is an appropriate solution for dealing with security issues associated with remote telephone access to a wide variety of telephone based applications like telephone banking, fund transfer, payment and transaction authentication, credit card authorization, telephone trading, voice dialing, accessing customer care services, password reset system etc [7]. With the help of an ASV system users can access any privileged information or services from anywhere in the world. Since dysarthric people are often physically debilitated and find it difficult to utilize user interfaces like a keyboard or a computer, such speech processing applications can be extremely beneficial in assisting dysarthric people. Through applications in employment, education, forensics, authentication and security, automatic speaker recognition technology can considerably improve the quality of life of the people suffering from dysarthria [8]. So, it should be given sufficient consideration and efforts in this field are essential to build an effective and robust ASV system for dysarthria patients.

Currently, majority of research related to dysarthric speech is primarily focused on automatic speech recognition [9], speech intelligibility enhancement [10], and automatic assessment [11]. On the other hand, only a few studies have attempted to address the issues of developing automatic speaker recognition systems for the people suffering from dysarthria. K L. Kadi *et al.* [11], proposed a speaker identification system for dysarthric speakers. In that work, the Mel-frequency cepstral coefficients (MFCC) were combined with acoustic features based on auditory cues. In addition to that, traditional classification methods such as Gaussian mixture models (GMM) and support vector machines (SVM) were employed for the identification task. A speaker identification system employing deep belief network for signal representation in combination with a classifier based on neural networks was proposed by Farhadipur *et al.* [1]. Feature representations including i -vectors, bottle-neck-neural-network-based features and covariance-based feature representation in combination with a multi-class SVM classifier were investigated for the task of dysarthric speaker recognition in [12]. However, it is worth emphasising here that automatic speaker verification is an unexplored area in the context of dysarthric speakers. Therefore, through the study reported in this paper, we presents our efforts in that direction.

The x -vector-based speaker embeddings extracted using a time-delay neural network (TDNN) are the current state-of-the-art representation employed for speaker verification task [13]. The work reported in this paper is the first attempt exploring the effectiveness of x -vector-based speaker embeddings while developing an ASV system for dysarthric speakers. It is to note that, the TDNN architecture employed in x -vector extraction necessitates that a large amount of domain-specific data is

used while training so as to robustly estimate the model parameters. Unfortunately, the amount of speech data from speakers suffering from dysarthria is very limited. Another important aspect of this study is how to deal with the paucity of domain-specific speech data in order to develop effective and robust ASV systems for dysarthric speakers. Several characteristics of dysarthric speech contribute in degrading the performance of the ASV systems. Speech-rate and average phoneme duration, are among the major factors reported in the literature. Motivated by these observations, we have explored the role of duration-modification-based data augmentation in order to increase the amount of domain-specific data. In this regard, the duration of speech data from healthy speakers is extended using an explicit signal processing technique. The modified speech is then pooled into training. As a consequence of duration-modification-based data augmentation, some of the relevant missing acoustic attributes are introduced into the training dataset. This in, turn, improves the accuracy of ASV system for dysarthric speakers.

The remainder of the paper is laid out as follows: Duration modification based data augmentation method is explained in Section 2. Experimental setup and results are presented in Sections 3. Finally, Section 4 concludes this paper.

2. Duration modification based data augmentation

In this section, the proposed data augmentation approach based on duration modification is discussed.

2.1. Need for data augmentation

In this paper, we present a text-independent dysarthric speaker verification system employing x -vector-based speaker embedding. Since the TDNN architecture employed in an x -vector system consists of several layers [13], large amount of speech data is necessary to develop these systems in order to properly utilise machine learning methodologies for ASV. A dysarthric speaker, on the other hand, finds it difficult to speak for long periods of time due to muscular weakness and exhaustion. Therefore collecting dysarthric speech data is a daunting task, particularly from speakers with high severity dysarthria. Consequently, available dysarthric speech databases consists of limited amount of speech data from a small number of speakers.

Training an x -vector-based system using limited amount of dysarthric speech, as already stated, will lead to under-fitting. On the other hand, using a large amount of speech data from healthy speakers for training the TDNN, will make the ASV system biased towards control subjects. This, in turn, will lead to poor performance with respect to speakers suffering from dysarthria. To overcome this issue of data scarcity and diversity, as well as to increase the robustness of the model, we adopted data augmentation. Data augmentation is the process by which we apply certain transformations to the existing training data to create new synthetic training samples and then pool it with the original dataset. The primary objective is to increase the diversity of the acoustic conditions captured by the training data and to enhance the ability of the trained models to generalize.

2.2. Motivation for duration-modification-based data augmentation

Due to difficulty with tongue and lip movement, dysarthric speakers may speak more slowly compared to the normal peo-

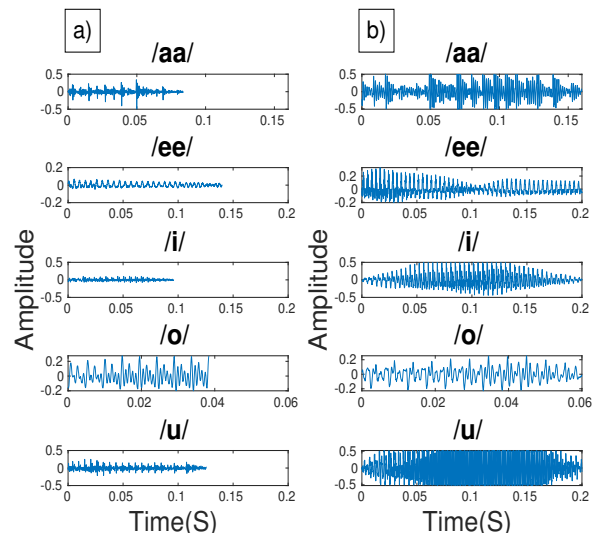


Figure 1: Waveform for vowel sounds /aa/, /ee/, /i/, /o/, /u/ spoken by (a) control subject, and (b) dysarthric subject

ple [14]. In order to get a better insight, we collected dysarthric and control speech and carried out some analysis using Matlab and Praat software. As an example, Figure. 1 shows the time domain waveform of vowel sounds (/aa/, /ee/, /i/, /o/, /u/) from control and dysarthric speech utterances. It is evident from the plot that vowel duration of dysarthric speech is longer than their control speech counterparts. This, in turn, implies that the total duration for the same set of sentences, spoken by dysarthric and control speakers, will be relatively longer in the case of dysarthric speakers. This is due to the inter word delays, frequent pauses, non-speech sounds and elongation of phonemes [15]. In addition to that, the average phoneme duration is also proportional to the degree of dysarthric speech severity, average phoneme duration increases with severity of the condition.

As stated earlier, the primary objective of data augmentation is to introduce the missing targeted acoustic attribute into the training data. One of the missing attribute in the context of ASV system for dysarthric speakers trained on control data (the ASV task explored in this study), is the increased average phoneme duration in the case of dysarthric speech. Motivated by this, we propose to extend the duration of the training data from control speakers and then pool it into training. As a result, the ASV system will learn larger phoneme duration and eventually become more robust towards dysarthric patients. This idea has been implemented in this study and we have experimentally validated that duration-modification-based data augmentation significantly improves the performance over the baseline system with respect to dysarthric speakers.

2.3. Employed ASV system architecture

The ASV system architecture employed in this work is illustrated in Fig. 2. A duration-modification-based data augmentation module is incorporated in the front end of the ASV system. Data augmentation helps to increase the diversity of the acoustic conditions captured by the training data and to introduce the missing targeted attributes as discussed earlier. The duration modification approach indiscriminately increases the duration of all phonemes in the utterances. For duration modification, instants around glottal closure (GC) and glottal opening (GO)

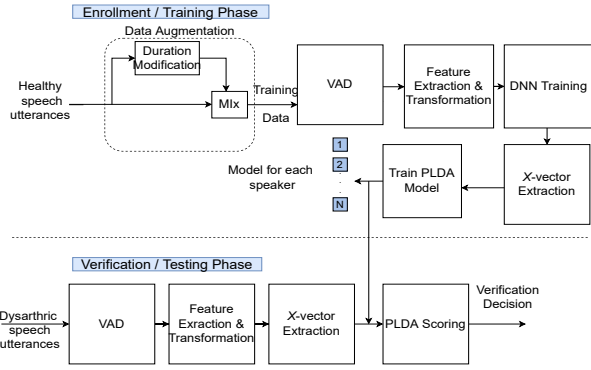


Figure 2: Simplified block diagram summarizing the system architecture employed in this work for verifying speech data from dysarthric speakers

are detected and exploited as described in [16]. It consist of two main tasks; 1) Compute the instants of glottal closure and opening, 2) Use those instants for duration modification. Zero-frequency filtering (ZFF) method is used for determining GC and GO instants and for finding the location of those instants.

2.3.1. ZFF method for computation of Glottal Closure and Opening Instant [17]

ZFF method for computation of GC and GO instant involves following steps:

1. Difference input speech signal $s[n]$,

$$x[n] = s[n] - s[n - 1] \quad (1)$$

2. Pass $x[n]$ twice through cascade of two ideal digital filter at zero frequency

$$r[n] = - \sum_{k=1}^4 b_k r[n - k] + x[n] \quad (2)$$

where $b_1=4, b_2=-6, b_3=4$ and $b_4=-1$ [17]

3. Remove the trend in filtered signal $r[n]$ by subtracting the average over 10 ms

$$\hat{r}[n] = r[n] - \frac{1}{2N+1} \sum_{n=-N}^N r[n + m] \quad (3)$$

where $2N+1$ corresponds to window size of average pitch period

Trend removed signal $\hat{r}[n]$ is the *Zero Frequency Filtered (ZFF) signal* and instants of GC and GO corresponds to *positive zero crossing of ZFF signal*.

2.3.2. Duration modification using GC and GO instant

Duration modification involves deriving new speech signal by including the desired modification in the duration of an utterance. It consist of three main tasks [18]:

1. Finding the GC and GO instant (epoch sequence) from the input speech signal
2. Derive a modified epoch sequence according to desired duration modification rate α

3. Reconstruct duration modified speech from modified GC and GO epoch sequence

α is the modification rate for duration modification of the training data. If $\alpha < 1$, it can result in time stretching, i.e., the duration of the reconstructed audio increases. If $\alpha > 1$, it can result in time compression, i.e., duration of the reconstructed audio decreases. $\alpha=1$ indicates that no modification have been made.

During the training phase, speech data from healthy speakers and its duration modified version are pooled together. The pooled training speech is passed through an energy-based voice activity detection (VAD) module to remove the non-speech sound units. Front-end feature extratcion is done next followed by feature normalization. Variable length utterances are then mapped to fixed-dimensional embedding called x -vectors using a TDNN trained on the pooled data. x -vector extraction is based on the system proposed in [13]. During evaluation or testing, feature extraction is performed on the speech data from dysarthric speakers. The TDNN-based extractor is then employed to obtain the corresponding x -vector. Finally, PLDA-based scoring is done for verification.

3. Experimental evaluation

3.1. Experimental setup

The dysarthric speaker verification system developed in this work is evaluated with two different dysarthric speech corpora namely Torgo database and Universal Access dysarthric speech (UA-Speech) corpus. Torgo database [19] consist of acoustic data from 8 dysarthric speakers (DS) with cerebral palsy (3 females and 5 males) with varying level of speech intelligibility and 7 control speakers (CS) (3 females and 4 males). On the other hand, the UA corpus [20] consists of utterances from 15 dysarthric speakers (4 females and 11 males) and 13 healthy control speakers (4 females and 9 males). These databases include speech intelligibility ratings for each dysarthric speaker, in terms of severity level. Based on the ratings, 23 dysarthric speakers in the current study were classified into 3 severity level categories namely low, medium and high. Utterances from control speakers were recorded under identical conditions and with the same vocabulary as that of dysarthric speakers, allowing us to compare the performance of ASV system for both dysarthric and normal speakers. Data from all the 23 dysarthric speakers of the Torgo and UA-Speech databases are used for evaluating the system performance. Two test sets were created using the data from Torgo and UA-speech corpora. The first test set referred to as Test-DS has a total of 230 utterances, 10 utterances from each of the dysarthric speakers in both databases. The second test set, Test-CS, contains a total of 200 utterances after considering 10 utterances from each of the control speakers (7 from Torgo and 13 from UA-Speech). The experiment was validated with a total of 52670 trials, 2070 genuine trials and 50600 impostor trials. For training purposes, the Vox-Celeb1 speech corpus [21], which consists primarily of data from healthy speakers, was utilised. This database consists of over 140k utterances from 1211 speakers.

ASV system development and evaluation were performed using the Kaldi speech recognition toolkit [22]. Equal Error Rate (EER) and minimum of normalized Detection Cost Function (DCF) at p -target=0.01 were used as the evaluation metrics.

3.2. Results and discussions

The ASV system trained exclusively using speech data from the Voxceleb1 database serves as the baseline. Baseline EER and

Table 1: EER and min DCF values for the baseline ASV system with respect to Test-DS and Test-CS test sets

	Test Set	
	Test-DS	Test-CS
EER (in %)	15.56	2.07
min DCF	0.675	0.250

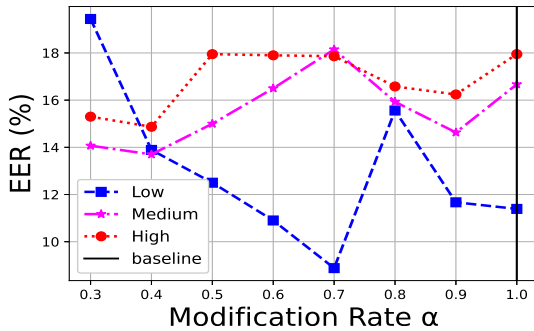


Figure 3: EER profile showing the impact of duration-modification-based data augmentation on dysarthric speaker verification task.

min DCF with respect to the two test sets are given in Table 1. As evident from the tabulated results, the baseline ASV system performs poorly in the case of dysarthric speakers. This is due to the stark differences in the acoustic attributes present in the training and test data as already discussed. To overcome this issue, duration-modification-based data augmentation was performed. For that purpose, we extended the duration of the training data from the VoxCeleb1 database by a modification rate α ($\alpha < 1$). α was varied from 0.3 to 0.9 in steps of 0.1. Decreasing α beyond 0.3 will result in the reconstruction of a training set that is too long in duration. This may affect the performance of low severity dysarthric speech test set. Modified speech data corresponding to each value of α was used as training data to construct distinct ASV systems. Performances of each of those ASV systems was evaluated separately using low, medium and high severity dysarthric speech test set. The variation of EER with change in α is shown in Figure 3.

It is to note that increasing the phone duration of training speech improves the performance of high and medium severity dysarthric speech while degrades the performance of low severity dysarthric speech. Based on observation, optimal values of modification factor chosen was 0.3, 0.4 and 0.8. Data obtained using these three scaling factors were all merged into training at the same and a final ASV system was trained. Training set of baseline ASV system is the original Voxceleb1 dataset and training set of proposed system is the Voxceleb1 dataset augmented with its duration modified versions at modification rates 0.3, 0.4 and 0.8. This proposed approach of duration-modification-based data augmentation is found to be effective and the same is evident from the EER and min DCF values with respect to Test-DS test set given in Table 2. Proposed system yielded much better performance for both healthy speakers as well as dysarthric speakers. Finally, EER and min DCF values were computed considering data from each of the severity levels in order to study the the impact of duration-

Table 2: EER and min DCF values for both test sets demonstrate the efficacy of proposed duration-modification-based data augmentation

ASV System	Test set	EER (%)	min DCF
Baseline	Test-DS	15.56	0.675
	Test-CS	2.07	0.250
Proposed	Test-DS	12.17	0.655
	Test-CS	1.25	0.131

Table 3: Severity wise comparison of the proposed approach over baseline in terms of EER and min DCF

ASV System	Severity	EER (%)	min DCF
Baseline	Low	11.39	0.488
	Medium	16.67	0.614
	High	17.95	0.743
Proposed	Low	11.11	0.477
	Medium	13.52	0.496
	High	13.33	0.639

modification-based data augmentation and the same are given in Table 3. These evaluation results show that proposed ASV system yielded much better performances even when the speech data was severely impaired due to dysarthria.

Dysarthric speakers depending on the severity, speak more slowly than normal speakers due to the weakened muscles. Those differences in phone duration lead to certain degree of acoustic mismatch and hence duration modification helps [5]. Duration modification helps to improve the performance for each individual severity level. Higher improvement was observed for high severity level approximately 26% and 14% relative improvement over baseline for EER and min DCF respectively. Low and medium severity levels showed a relative improvement of 2.5% and 19% for EER and 2.3% and 19.3%, respectively for min DCF over the baseline system. Hence a single ASV system can be effectively used for verification of normal as well as dysarthric speaker of varying speech intelligibility, by suitably modifying the duration of the speech. To the best of our knowledge, current work is the first attempt to dysarthric speaker verification.

4. Conclusions

Speaker recognition is a challenging problem for dysarthric speakers. In this paper we proposed an automatic speaker verification (ASV) system for dysarthric speakers with varying speech intelligibility. To improve the performance of the baseline ASV system with respect to dysarthric speakers, we incorporated a duration-modification-based data augmentation module in the front end of the ASV system. Duration modification with several scaling factors was applied to the healthy training data and the modified data was then augmented with the original version for training the ASV system. The experimental results demonstrate the efficiency of duration modification, which resulted in a relative improvement of 22% over the baseline. Proposed method helps to improve the performance for each individual severity level as well. Approximately 26% relative improvement over the baseline was noted for high severity level.

5. References

- [1] A. Farhadipour, H. Veisi, M. Asgari, and M. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *ETRI Journal*, vol. 40, 07 2018.
- [2] J. Barkmeier-Kraemer and H. Clark, "Speech–language pathology evaluation and management of hyperkinetic disorders affecting speech and swallowing function," *Tremor and Other Hyperkinetic Movements*, vol. 7, Sep. 2017, publisher Copyright: © 2017 Barkmeier-Kraemer et al.
- [3] "Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34, no. 3, pp. 309–316, 1997.
- [4] L. Pennington, N. K. Parker, H. Kelly, and N. Miller, "Speech therapy for children with dysarthria acquired before three years of age," *The Cochrane database of systematic reviews*, vol. 7, p. CD006937, 2016.
- [5] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," 09 2018, pp. 471–475.
- [6] M. N. Nachappa, A. M. Bojamma, C. Prasad, and M. Nithya, "Automatic speaker verification system," 2014.
- [7] R. Naika, "An overview of automatic speaker verification system," in *Intelligent Computing and Information and Communication*, S. Bhalla, V. Bhateja, A. A. Chandavale, A. S. Hiwale, and S. C. Satapathy, Eds. Singapore: Springer Singapore, 2018, pp. 603–610.
- [8] M. Chaiani, M. Bengherabi, S. A. Selouani, and M. Boudraa, "Dysarthric speaker identification with constrained training durations," in *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, 2018, pp. 1–6.
- [9] J. Ren and M. Liu, "An automatic dysarthric speech recognition approach using deep neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 8, 01 2017.
- [10] M. S. Yakoub, S.-A. Selouani, and D. O'Shaughnessy, "Improving dysarthric speech intelligibility through re-synthesized and grafted units," in *2008 Canadian Conference on Electrical and Computer Engineering*, 2008, pp. 001 523–001 526.
- [11] K. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, 11 2015.
- [12] M. Senoussaoui, M. Sarria-Paja, P. Cardinal, T. Falk, and F. Michaud, *1. State-of-the-art speaker recognition methods applied to speakers with dysarthria*, 02 2020, pp. 7–34.
- [13] 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)*, 2018, pp. 5329–5333.
- [14] Y. Yunusova, G. Weismer, J. Westbury, and M. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of speech, language, and hearing research : JSLHR*, vol. 51, pp. 596–611, 07 2008.
- [15] A. Prakash, M. Reddy, and H. Murthy, "Improvement of continuous dysarthric speech quality," 09 2016, pp. 43–49.
- [16] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, "Effect of prosody modification on children's asr," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1749–1753, 2017.
- [17] S. R. M. Prasanna, D. Govind, K. S. Rao, and B. Yegnanarayana, "Fast prosody modification using instants of significant excitation," 2010.
- [18] K. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.
- [19] F. Rudzicz, A. Namasivayam, and T. Wolff, "The toro database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 1–19, 01 2010.
- [20] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," 01 2008, pp. 1741–1744.
- [21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," 06 2017.
- [22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesel, "The kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.