Interspeech 2018
2-6 September 2018, Hyderabad

Multimodal Speaker Segmentation and Diarization using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks

Tae Jin Park, Panayiotis Georgiou

University of Southern California, Los Angeles, CA, USA

taejinpa@usc.edu, georgiou@sipi.usc.edu

Abstract

While there has been substantial amount of work in speaker diarization recently, there are few efforts in jointly employing lexical and acoustic information for speaker segmentation. Towards that, we investigate a speaker diarization system using a sequence-to-sequence neural network trained on both lexical and acoustic features. We also propose a loss function that allows for selecting not only the speaker change points but also the best speaker at any time by allowing for different speaker groupings. We incorporate Mel Frequency Cepstral Coefficients (MFCC) as an acoustic feature stream alongside lexical information that are obtained from conversations from the Fisher dataset. Thus, we show that acoustics provide complementary information to the lexical modality. The experimental results show that sequence-to-sequence system trained on both word sequences and MFCC can improve on speaker diarization result compared to the system that only relies on lexical modality or the baseline MFCC-based system. In addition, we test the performance of our proposed method with Automatic Speech Recognition (ASR) transcripts. While the performance on ASR transcripts drops, the Diarization Error Rate (DER) of our proposed method still outperforms the traditional method based on Bayesian Information Criterion (BIC).

Index Terms: Speaker Diarization, Speaker Segmentation, Sequence to Sequence Models

1. Introduction

Speaker Diarization is an important pre-processing step towards a complete Automatic Speech Recognition (ASR) system that includes multiple speakers. Further, speaker diarization information plays a crucial role in speech analytics such as turn-taking characteristics and is critical in many behavioral analytics applications [1, 2]. Poor performance of speaker diarization is bound to deteriorate the performance of subsequent models such as ASR, emotion recognition, behavioral informatics, and topic analysis systems. Speaker segmentation is a critical component of this process and heavily affects the performance of speaker diarization and hence all subsequent modules.

In general, a speaker diarization system consists of two main parts: segmentation and clustering. Segmentation aims to detect all speaker change points. The most widely used method is the Bayesian Information Criterion (BIC) based segmentation [3, 4]. More recently, methods based on Recursive Neural Networks (RNN) have shown improved performance on speaker segmentation [5, 6]. In addition, Joint Factor Analysis (JFA) [7] has also shown promising results. Further, there are significant efforts in speaker segmentation and diarization with pre-trained Deep Neural Networks (DNN) both through supervised-training [8] and through unsupervised-training [9, 10].

Despite the very active field, there has been very little effort in exploiting lexical information towards this task. Most of the research that involves lexical information or transcript is relating to speaker identity [11, 12] or speaker role [13, 14]. India *et al.* employed character level information via an LSTM network with a character level Convolutional Neural Network (CNN) and i-vector training on transcript [15].

One likely reason that transcripts from ASR have not been used for diarization is that we often are hesitant to run ASR before diarization since that will be more noisy that employing these two components in reverse order. However that is not a constraint (except in computation resources) as the ASR can be re-run after diarization a second time. Further, along recent efforts of research including in our group, of joint training, future implementations can jointly optimize for diarization and ASR.

In this work, we propose a system that incorporates both lexical cues and acoustic cues to build a system closer to how humans employ information. We investigate a sequence-to-sequence model (seq2seq) that integrates both lexical and acoustic cues to perform speaker segmentation and speaker diarization. Sequence-to-sequence models have been widely used for language translation [16], end to end ASR systems [17] and text summarization[18]. The advantage of seq2seq over Recurrent Neural Network (RNN) based models (LSTM [19], GRU [20]) is that it can summarize the whole sequence into an embedding and then pass it to the decoder. Moreover, it can integrate information and process variable length sequences. In doing so, such a model can capture temporally encoded information from both before and after the speaker change points. In addition, the attention mechanism of this model helps in capturing the important parts of characterizing the speaker(s).

In our work we employ dyadic-interaction data to train and test the proposed system. Our proposed model operates on both reference transcript data and, critically for realistic deployments, on ASR hypotheses.

2. Proposed Speaker Diarization System

2.1. Network Architecture

Our proposed sequence to sequence model consists of encoder, decoder and attention model that connects encoder and decoder. The encoder consumes a sequence of word representations, along with acoustic features (MFCC) described in sec. 2.2, as shown in Fig. 1. The decoder produces a sequence of words along with speaker IDs during the speaker change points, as shown in Fig. 2. We use GRU with a 256-dimensional hidden layer and an attention model that has been applied to many state-of-the-art machine translation systems [21].

2.2. Feature processing

In our proposed method the features are time-synchronous. All the features align with the word boundaries as follows:

WORD: The word sequences we use are obtained either from

10.21437/Interspeech.2018-1364

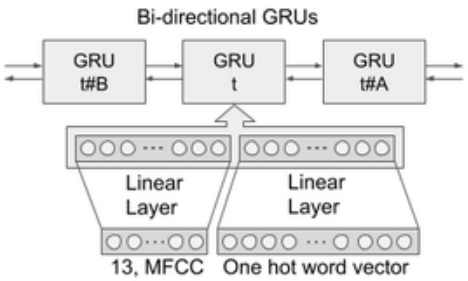

Figure 1: *The encoder side of the proposed network.*

Table 1: *An example of source sentence and target sentence in training data.*

Source	hello hi my name is James hi James
Target	hello ♯A hi ♯B my name is James ♯A hi James

the reference transcripts or from ASR outputs. We use a linear layer to convert one-hot word vector into word embedding as described in Fig. 1. The source sequence is 32 words in the reference transcripts or ASR outputs. The target sequence for training is 32 words and added speaker turn tokens as in the example sentence in table1.

MFCC: We use 13-dimensional MFCCs extracted with a 25ms window and 10ms shift. Detailed specifications follow the default settings in [22]. We then average the MFCC features for the word-segment and thus derive a 13×1 vector for each word.

2.3. Encoder and input features

In our proposed system, the encoder integrates MFCC feature vectors and word embeddings. Fig.1 shows how the proposed encoder is structured. Word embeddings and MFCC features are connected through linear layers. After the fully-connected layers, the embeddings are concatenated. The concatenated vector is then fed to the GRU that is the encoder of the seq2seq system. We use 256 hidden unit size, word embedding size and output layer of linear layer for MFCC vector. The number of hidden layers were chosen to be equal for both MFCC and word embedding because there is a performance degradation when these embedding size are different. However, more optimization needs to take place for the optimal system.

2.4. Decoder and loss function

In our proposed system, the decoder outputs a word sequence and the speaker turn token "♯A" and "♯B". Fig. 2 describes the decoder side in our proposed system. Unlike word tokens, the loss of the speaker turn tokens are calculated in a different way that ignores the speaker IDs and only focuses on speaker groupings. For example, the speaker turn sequence of "♯A ♯B ♯A" is considered equal to "♯B ♯A ♯B". That is, the loss function in our proposed system calculates two versions of losses: original and flipped version of speaker turn tokens. Between these two losses, our loss function selects the smaller loss. This loss function also avoids learning the probability between speaker turn tokens and words in the target sequences in the training set.

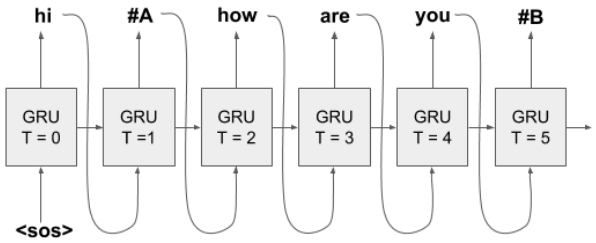

Figure 2: *The decoder side of the proposed sequence to sequence model.*

2.5. Speaker Turn Estimation

To maximize the accuracy of speaker turn detection, we employ shift and overlap scheme to predict the speaker turn. Fig. 3 explains how speaker turn prediction is done. A target window that has 32 word length sweeps the whole session from the beginning to the end. For each target window, we predict speaker turn tokens with our trained sequence to sequence model. At each prediction, we extract 32 words and 32 MFCC vectors from transcripts and audio signals, respectively. A set of speaker turns for a session is estimated through the following process in accordance with the indices in Fig, 3.

1. Obtain a new word sequence and estimated speaker turn tokens from decoder outputs.

2. Form a speaker turn vector by assigning each word the nearest speaker turn token.

3. Store the speaker turn vector that is obtained from step 2 in a cumulative speaker turn sequence which is the matrix that sequentially stores all the speaker turn vectors obtained so far. Flip the speaker turn vector if flipping the speaker turn vector gives less hamming distance with all the other speaker turn tokens in cumulative speaker turn sequence.

4. Store the speaker turn vector from step 3 into the cumulative speaker turn sequence. Shift one word to the right and feed next 32 words and 32 MFCC vectors to the encoder of the proposed system.

After finishing the above process by shifting 32 word window to the end of the session, we determine the final speaker turn decision by taking a majority vote. In this way, a word in a session incorporates 32 different predictions to determine the speaker turn.

2.6. Clustering

We will evaluate on diarization accuracy we therefore employ our SCUBA, BIC based agglomerative clustering algorithm based on [4] to perform the clustering step. For the agglomerative clustering we employ the raw frame-level MFCC as features. We obtain the segmented MFCC streams using speaker turn information that is produced from the process described in 2.5. This clustering algorithm is applied to all of the models in this paper, including the LIUM baseline. For the baseline systems, the process mentioned in 2.5 is replaced with other methods while same agglomerative clustering algorithm is applied.

3. Experimental Results

Our proposed system is tested with two different datasets: those stemming from reference transcription and those from automat-

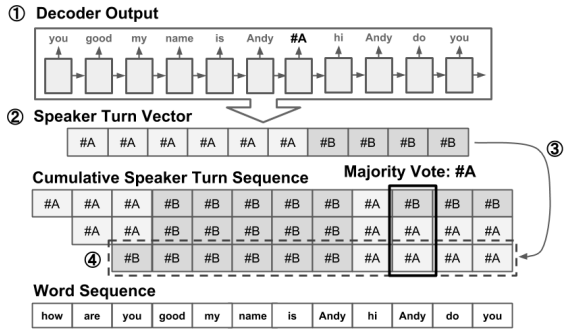

Figure 3: *Decoder outputs and overlapping speaker turn vectors.*

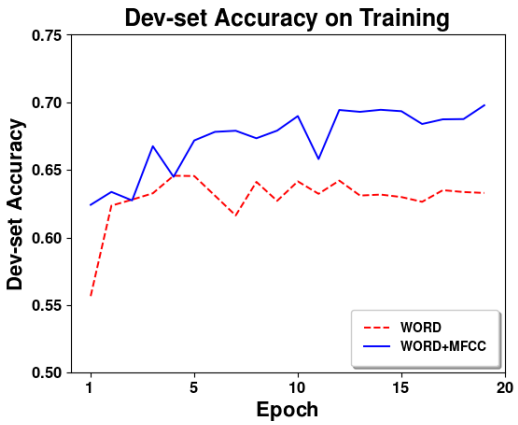

Figure 4: *Dev-set accuracy on training.*

ically derived ASR hypotheses.

To train our proposed system with dialogue, we train our proposed system on Fisher English Training Speech Part 1 and Part 2 [23] for both lexical cues and acoustic cues. This results in 11,112 training dialogs comprised of approximately 19 million words.

Before training the proposed system, we randomly chose and separated 20 sessions as a test set and 567 sessions as a dev-set from the original Fisher dataset. These are used as evaluation in the case we employ clean transcripts. For evaluation using ASR outputs, we also use Switchboard-1 Telephone Speech Corpus [24] to ensure complete train-test separation and domain generalization. Although the original recordings were 2-channel telephony (1 per speaker) we generate single channel signals by mixing down to mono. For the word alignment information, we use forced-alignment to obtain the word alignment information for Fisher dataset since word-level alignment information is not provided in Fisher dataset while speaker turn level alignment is provided. For Switchboard-1 dataset, we use the provided word alignment information and speaker turn level alignment information. With this alignment information, we create the ground truth diarization labels for subsequent evaluation. Due to the overlaps in the data the lower-bound diarization error is not zero, and we will thus also denote that in the tables below.

As a benchmark of our proposed method we employ LIUM Speaker Diarization Tools [25] which contains a Speaker Activity Detection (SAD) system and a speaker segmentation system. The LIUM script that we use performs MFCC feature extraction, SAD and speaker segmentation sequentially. We use default settings for all the parameters. The clustering step is employing the same algorithm as all other methods in this paper (*i.e.,* LIUM segmentation and SCUBA clustering)

A second baseline is to employ agglomerative clustering for diarization but by employing the word boundaries as segmentation. For convenience, we refer to this model as WS. WS baseline can verify the merit of our proposed model since we can compare whether the performance is stemming from word alignment or speaker turn probability when we estimate with our proposed system. For reference transcript based test, WS is obtained from word alignment data in the transcript and for ASR transcript based test, WS is obtained from word alignment data from ASR transcript. We are using Diarization Error Rate (DER) as a performance metric for all experiments. To measure the DER metric, we employ the *md-eval* software in RT06S dataset [26] with the forgiveness collar of 0.25 seconds.

3.1. Training of sequence to sequence model

We train and test two different models separately. Each model employs the same architecture and the same training conditions except the feature types. The first model is trained only on word embeddings while the second model is trained on both word embeddings and MFCC. For convenience, we will refer to these as W model and WM model respectively. We train each model until convergence (20 epochs). We use teacher forcing [27] ratio of 0.5 to speedup training. Fig. 4 shows the dev-set accuracy while training. The WM model clearly shows improved performance over W model. Note that accuracy in Fig. 4 is accuracy measured with word sequence that contains speaker turn tokens and word tokens. Thus, this accuracy does not always mean better segmentation or diarization accuracy.

3.2. Experiment on Reference Transcripts

First, we do an experiment using reference transcripts. In this case MFCC features are obtained using the oracle word alignments. Thus, we use accurate word embedding and temporal information of each word. Table 2 shows the results we obtained from transcript data.

The result clearly shows that incorporating MFCC features helps the performance of diarization when the word embeddings and temporal information are accurate. In addition, W model and WM model also outperformed word-level segmentation (WS) based result. This suggests that applying our proposed model gives a merit over simply using word-alignment information as segmentation result. We also test the diarization system with ground truth speaker label per word and it shows the accuracies of 16.22% and 18.06% for Fisher and Switchboard data respectively. This is due to the frequent overlaps in dialogues and inaccurate labeling of speaker turn level transcript data. Therefore, "Oracle" DER in table 2 is the best performance we can achieve with any algorithm. To check the performance of the proposed system in different way, we also measure Word-level Diarization Error Rate (WDER) which means "who says this word". Table 3 shows WDER result for transcript based experiment. Since there are two speakers in this experiment, the WDER also shows similar result to the DER result where WM model shows nearly 4% improvement over W model.

Table 2: *DER on transcription data.*

DER(%)	W	WM	WS	Oracle	LIUM
Fisher	28.02	24.26	44.53	16.22	77.45
Switchboard	27.89	22.44	46.4	18.06	66.57

Table 3: *WDER on transcription data.*

WDER(%)	W	WM
Fisher Transcript	16.42	12.32
Switchboard Transcript	12.4	8.56

3.3. Experiment on ASR transcript

For ASR transcripts, we use the Kaldi Speech Recognition Toolkit [28] and ASR model trained on whole Fisher English Speech data. As a test-set, we choose the 30 audio files that have lowest index in each of 30 folders in Switchboard-1 dataset for reproducibility of our experiment. Table 4 shows the results from ASR based experiment. Unlike in the case of reference transcripts, WM model did not improve the performance. However, ASR based results are still better than diarization based on segmentation result obtained from LIUM Speaker Diarization Tools. In addition, WS model also performs better than LIUM Speaker Diarization Tools, which indicates using word-level segmentation from ASR can still perform better than BIC based segmentation system.

3.4. WER vs DER

Since we test the improvement by incorporating acoustic cues with transcript data, performance degradation in the experiment with ASR transcript is solely caused by poor ASR Word Error Rate (WER). The average WER for 30 Switchboard session is 35.15%. Fig. 5 shows the scatter plot between WER vs DER for the experiment with ASR transcript (Table 4). As we can see in Fig. 5, no session shows low DER when WER is high. However, although WER is pretty low, DER can be very high. Based on this outcome, we could conclude that low WER is necessary condition for low DER, not the sufficient condition.

4. Discussion

Comparing the two experiments using the reference transcripts and ASR transcripts with our proposed system shows that ASR performance hugely affects the performance of DER. However, the experiment with transcript still shows that acoustic cues can improve the diarization performance. Therefore, we can conclude that acoustic cues can be integrated with lexical cues but the ASR performance is critical. Further we believe that many of the errors that are made by the ASR in the segmentation step may create unrecoverable errors, and hence this points to potential benefits of using lattice information and exploiting the ASR uncertainty.

5. Conclusions

In this paper, we investigated the way to integrate lexical cues and acoustic cues with sequence to sequence model to improve speaker diarization performance. The results show strong support that lexical information can improve the speaker diarization system. We also see that ASR performance plays a crucial role

Table 4: *DER on ASR transcript and baseline system.*

DER(%)	W	WM	WS	Oracle	LIUM
Switchboard ASR	38.64	50.95	46.02	18.06	66.57

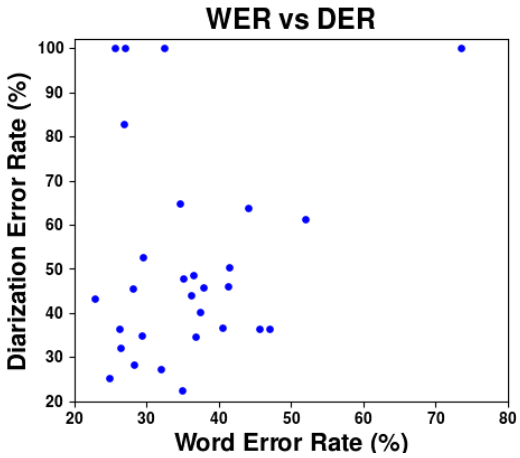

Figure 5: *Scatter plot of WER vs DER*

to the performance of our proposed system and poor WER degrades the proposed system trained on both acoustic features and word embeddings. The future work might include improving performance by training data on ASR transcript including multiple-hypotheses to provide alternate word alignment and segmentation points. Further we will investigate the use of alternate acoustic feature representations such as i-vector or embeddings obtained from neural networks[10, 9]. In addition, a fusion of frame and word level segmentation will also be considered to increase flexibility on segmentation decisions.

6. Acknowledgements

The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

7. References

[1] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. Scottsdale, AZ: ACM, 2011, pp. 7–12.

[2] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1 –31, 2013.

[3] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[4] S. Chen, P. Gopalakrishnan *et al.*, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[5] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Proc. Interspeech 2017*, 2017, pp. 3827–3831.

[6] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," *arXiv preprint arXiv:1710.10468*, 2017.

[7] B. Desplanques, K. Demuynck, and J.-P. Martens, "Factor analysis for speaker segmentation and improved speaker diarization," in *16th Annual conference of the International Speech Communication Association (INTERSPEECH 2015)*. International Speech Communication Association (ISCA), 2015, pp. 3081–3085.

[8] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4930–4934.

[9] A. Jati and P. Georgiou, "Speaker2vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation," *Proc. Interspeech 2017*, pp. 3567–3571, 2017.

[10] ——, "Neural predictive coding using convolutional neural networks towards unsupervised learning of speaker characteristics," *IEEE Trans. Speech, Audio, and Language Processing*, 2018.

[11] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts." ICSLP, 2004.

[12] Y. Esteve, S. Meignier, P. Deléglise, and J. Mauclair, "Extracting true speaker identities from transcriptions," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[13] B. Xiao, P. Georgiou, Z. E. Imel, D. Atkins, and S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PLOS ONE*, December 2015.

[14] B. Xiao, C. Huang, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling," *PeerJ Computer Science*, vol. 2, p. e59, Apr. 2016.

[15] M. À. India Massana, J. A. Rodríguez Fonollosa, and F. J. Hernando Pericás, "Lstm neural network-based speaker segmentation using acoustic and language modelling," in *INTERSPEECH 2017: 20-24 August 2017: Stockholm*. International Speech Communication Association (ISCA), 2017, pp. 2834–2838.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

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

[18] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] J. Lyons, "Python speech features," https://github.com/jameslyons/python-speech-features, 2017, accessed: 2018-03-23.

[23] C. Cieri, D. Miller, and K. Walker, "Fisher english training speech parts 1 and 2," *Philadelphia: Linguistic Data Consortium*, 2004.

[24] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium, Philadelphia*, vol. 926, p. 927, 1997.

[25] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Interspeech*, 2013.

[26] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[27] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

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