INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Articulatory Feature Extraction using CTC to build Articulatory Classifiers Without Forced Frame Alignments for Speech Recognition

Basil Abraham, S. Umesh, Neethu Mariam Joy

Indian Institute of Technology-Madras, India
{ee11d032,umeshs,ee11d009}@ee.iitm.ac.in

Abstract

Articulatory features provide robustness to speaker and environment variability by incorporating speech production knowledge. Pseudo articulatory features are a way of extracting articulatory features using articulatory classifiers trained from speech data. One of the major problems faced in building articulatory classifiers is the requirement of speech data aligned in terms of articulatory feature values at frame level. Manually aligning data at frame level is a tedious task and alignments obtained from the phone alignments using phone-to-articulatory feature mapping are prone to errors. In this paper, a technique using connectionist temporal classification (CTC) criterion to train an articulatory classifier using bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) is proposed. The CTC criterion eliminates the need for forced frame level alignments. Articulatory classifiers were also built using different neural network architectures like deep neural networks (DNN), convolutional neural network (CNN) and BLSTM with frame level alignments and were compared to the proposed approach of using CTC. Among the different architectures, articulatory features extracted using articulatory classifiers built with BLSTM gave better recognition performance. Further, the proposed approach of BLSTM with CTC gave the best overall performance on both SVitchboard (6 hours) and Switchboard 33 hours data set.

Index Terms: ASR, articulatory features, CTC, BLSTM

1. Introduction

Automatic speech recognition (ASR) systems have seen significant improvement in performance with recent advances in deep neural network architectures. However, factors like speaker variability and environment noise continue to degrade the performance. Hence, developing techniques to generate features or models that are robust to these effects are very important. Articulatory features (AF) are a way of incorporating speech production knowledge into ASR [1]. It has been reported in [2, 3, 4, 5, 6] that the use of speech production knowledge makes the models more robust to speaker and channel variations. Articulatory features can be generated [7] either by directly measuring the articulatory parameters using cine-radiography, or by performing inverse filtering of acoustic signal, or by using articulatory class probabilities obtained from articulatory classifiers built from speech data. Articulatory features obtained from articulatory classifiers are often referred to as pseudo articulatory features. Our work is restricted to the use of pseudo articulatory features and from now on we will refer to them as only articulatory features.

The use of hybrid artificial neural network for articulatory features extraction was studied in [7, 8, 9]. In [7], articulatory features in adverse environmental conditions were studied and shown to give promising results. The details and results obtained in Johns Hopkins 2006 summer workshop on Articulatory Feature-based Speech Recognition are given in [8, 9]. The use of articulatory features in tandem HMM paradigm are described in the report.

To extract pseudo-articulatory features in any language, AF classifiers need to be constructed for each of the AF group (eg: Place of articulation, Degree and Manner of articulation). To build a robust AF classifier, large amounts of data transcribed in terms of AF values (eg: alveolar, dental) in that language is required. To get the data aligned at frame level, manual transcription techniques were employed in [10]. It is difficult to manually transcribe data at frame-level in terms of AF values. Hence, the usual practice is to obtain a phone-level alignment and convert it into AF values using the phone-to-AF mapping.

In this paper a novel technique using connectionist temporal classification (CTC) is proposed. The proposed technique eliminates the need of forced frame level alignment of articulatory values. CTC [11] is a sequence labeling technique which was recently proposed and used for speech recognition in [11, 12, 13]. When compared to existing techniques, our proposed method provides an absolute improvement of 9% on SVitchboard task [14] and 4% on Switchboard 33 hours task [15].

In this paper various neural network architectures are used to extract AF and their performances are analyzed. The Neural network architectures studied include deep neural networks (DNN), convolutional neural network (CNN) and bidirectional long short-term memory (BLSTM) recurrent neural network (RNN). Among these architectures BLSTM gave better performance and along with CTC gave the best performance in both SVitchboard and Switchboard tasks.

The paper is organized as follows. Sections 2 and 3 describes articulatory feature set and CTC criterion respectively. The proposed articulatory feature extraction technique is explained and analyzed in section 4. Section 5 compares articulatory classifiers built with and without alignments. Articulatory features are further analyzed in section 6 and conclusions are drawn in section 7.

2. Articulatory Features

Pseudo-articulatory features (pseudo-AF) are used in our work to build acoustic models. The use of these features for ASR were described in [7, 9, 8]. In [9] a discrete multi-level feature set was introduced. The feature set has eight articulatory groups and their details are given in Table 1. In the case of diphthongs, (eg. /aw/), the initial state is denoted by /aw1/ and final state by /aw2/. A 'silence' class is also associated with each of the articulatory group.

http://dx.doi.org/10.21437/Interspeech.2016-925

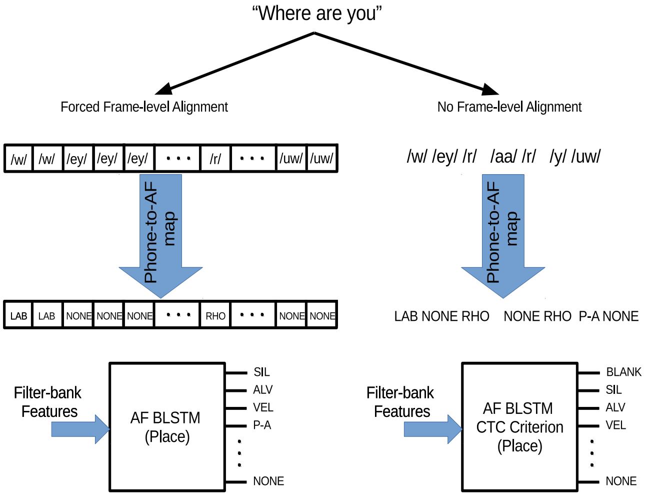
"Where are you"
Forced Frame-level Alignment
No Frame-level Alignment
/w/ /w/ /ey/ /ey/ /ey/ • • • /r/ • • • /uw/ /uw/
/w/ /ey/ /r/ /aa/ /r/ /y/ /uw/
Phone-to-AF map
Phone-to-AF map
LAB LAB NONE NONE NONE • • • RHO • • • NONE NONE
LAB NONE RHO NONE RHO P-A NONE
Filter-bank Features
AF BLSTM (Place)
SIL
ALV
VEL
P-A
NONE
Filter-bank Features
AF BLSTM CTC Criterion (Place)
BLANK
SIL
ALV
VEL
NONE

Figure 1: Building "Place" Articulatory Classifier

Table 1: Articulatory feature set

Group	Card-inality	Feature values
Place	10	alveolar (ALV), dental (DEN), labial (LAB), labio-dental (L-D), lateral (LAT), none, post-alveolar (P-A), rhotic (RHO), velar (VEL)
Degree & Manner	6	approximant (APP), closure (CLO), FLAP, fricative (FRIC), vowel (VOW)
Nasality	3	-, +
Rounding	3	-, +
Glottal State	4	aspirated (ASP), voiceless (VL), voiced (VOI)
Vowel shape	23	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, nil
Height	8	HIGH, LOW, MID, mid-high (MID-H), mid-low (MID-L), very-high (VI), nil
Frontness	7	back (BK), front (FRT), MID, mid-back (MID-B), mid-front (MID-F), nil

3. Connectionist Temporal Classifier (CTC)

CTC is a sequence labeling technique recently introduced for speech recognition [11, 12, 13]. It tries to find the best alignment between the input speech frames and output labels. The main concept behind CTC criterion is to interpret the network outputs as a probability distribution over all possible label sequences for a given input feature set. CTC uses a softmax layer to give posterior probabilities to all output symbols at each time step. In ASR, CTC uses phones as the output symbols along with an extra blank symbol to denote no-label at any time step. This is followed by many-to-one mapping which maps output labels to a label sequence without blank. This mapping removes blanks and any repeated labels in the output label sequence. CTC uses a forward-backward algorithm to sum over all possible alignments of output labels which can represent the target sequence.

4. Proposed Feature Extraction Technique

In this section, the steps involved in extracting articulatory features using articulatory classifiers built with CTC criterion is explained. These articulatory classifiers need to be trained for each of the AF group as given in Table 1. In previous works [8, 9], the articulatory classifiers were built using frame-level alignment of AF values corresponding to that AF group. In those works the frame-level alignment of AF values were generated manually using the procedures given in [10]. Another way to generate frame level AF value alignments is by converting the frame-level phone alignments using a phone-to-articulatory feature mapping in that language [7].

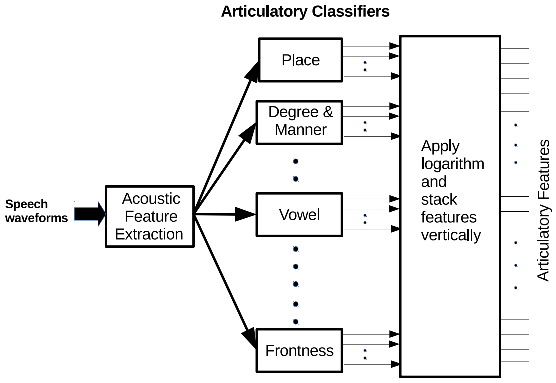

Figure 2: Articulatory Feature Extractor

The efficacy of the articulatory features extracted depends on the quality of the articulatory classifiers. Hence the articulatory classifiers need to be built using error free frame-level alignments. Generating error free alignment manually is a difficult task. In this paper, a technique to generate articulatory classifiers without forced-frame level alignments is proposed. The proposed technique uses CTC criterion to build a BLSTM articulatory classifier. The CTC criterion avoids the requirement of frame alignments. The steps involved in proposed method are as follows:

Algorithm 1: Steps Involved in Articulatory Feature Extraction

1. Convert the train transcription in terms of phones to AF values using phone-to-AF mapping in the language as shown in Figure 1.
2. Train BLSTM classifiers with CTC criterion from transcription obtained in step 1 with AF values and an extra blank symbol (BLANK).
3. For both train and test data, extract the posterior features from each of the AF-BLSTM articulatory classifier.
4. Apply logarithm to expand the dynamic range as shown in Figure 2.
5. Concatenate the output features from all the AF-BLSTM classifiers as shown in Figure 2 to get the articulatory features.

Figure 3 shows the comparison of AF value probabilities for the articulatory group "Place" obtained from BLSTM classifier built with CTC and frame-level alignment. The same utterance as shown in Figure 1 was used for the analysis. From Figure 3, it can be seen that the output labels of CTC classifier follows the target sequence closely compared to frame-level classifier. BLSTM classifier using frame alignments had many cases of misclassification compared to BLSTM-CTC. This supports our claim that articulatory classifier built with CTC performs superior to the ones trained on frame alignments. As seen in Figure 3, P-A is very poor when using frame alignments.

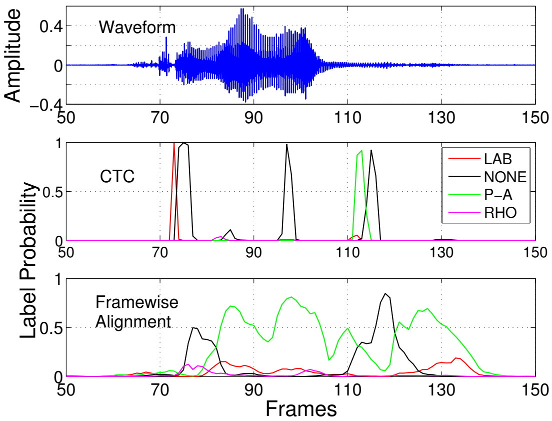

Figure 3: Analysis of "Place" Articulatory Features Extracted with CTC and Frame-level alignments

5. Comparison of various Networks for AF classifier using Frame Alignments and CTC

In this section, we perform a detailed analysis on the effectiveness of different neural network architectures for building articulatory classifiers. The analysis was performed for the architectures like DNN, CNN and BLSTM. The BLSTM was built using both, cross-entropy criterion with frame alignments and CTC criterion without any frame alignments. The DNN and CNN were built using frame level alignments. Articulatory classifiers using frame level alignments were built using Kaldi toolkit [16] and the BLSTM with CTC criterion was built with EESEN toolkit [17] based on Kaldi toolkit. The articulatory classifiers (for Degree, Place, etc.,) were built with 110 hours of Switchboard database. In each of articulatory classifiers the output targets are the corresponding AF values in that AF group. The configurations for the different neural network architectures are given below:

- **AF-DNN** was trained with 3 layers and 1024 nodes in each layer for Nasality, Rounding and Glottal classifiers and rest of the articulatory classifiers had 2048 nodes. AF-DNN was trained with restricted Boltzmann machine (RBM) pretraining followed by DNN training.
- **AF-CNN** was trained with 2 layers and 1024 nodes in each layer for all the classifiers. The CNN used a covolutional window of size 8. Pooling layer window was of size 3 and no overlapping of pooling window was allowed. The CNN layers had 128 feature maps in the first layer and 256 feature maps in the second layer.
- **AF-CNN-DNN** was trained with 2 layers of CNN followed by 4 layers of fully connected DNN layers. The CNN layers were trained in the same way as that in AF-CNN and the output layer was removed to generate the features for the DNN layers. The DNN layers were trained with 1024 nodes in each hidden layer.
- **AF-BLSTM** was trained with 2 layers and cell dimension of 256 for all articulatory classifiers.
- **AF-BLSTM-CTC** used the same configurations as that of AF-BLSTM and was trained using CTC objective function. AF values along with an extra blank symbol are used as output targets .

Table 2: Performance Comparison (%WER) of AF Extracted using Different Neural Network Architectures for AF Classifiers in SVitchboard Corpus

(a) Test Set

Acoustic Model Type	MFCC	Classifiers built with Alignments				CTC
		AF-DNN	AF-CNN	AF-CNN-DNN	AF-BLSTM	AF-BLSTM-CTC
Monophone	70.92	53.37	51.5	47.54	55.65	55.25
Triphone	56.46	44.64	43.41	43.67	41.35	42.78
LDA-MLLT	53.46	42.76	44.16	41.99	38.34	32
DNN	44.66	37.43	39.46	36.37	33.95	28.63

(b) Dev Set

Acoustic Model Type	MFCC	Classifiers built with Alignments				CTC
		AF-DNN	AF-CNN	AF-CNN-DNN	AF-BLSTM	AF-BLSTM-CTC
Monophone	70.39	53.17	50.9	47.6	54.88	55.07
Triphone	55.24	44.42	42.12	43.26	39.89	43.46
LDA-MLLT	52.21	41.61	43.03	41.24	37.46	32.54
DNN	43.53	36.74	38.99	36.12	33.31	29.03

Now we compare the various articulatory features obtained above in terms of recognition performance (word error rate (WER)). The next section discusses the experimental set-up and the databases used.

5.1. Databases

The SVitchboard task [14] is commonly used for articulatory feature related works. It is a small vocabulary task defined using the subsets of Switchboard-1 corpus with words ranging from 10 to 500. The database is divided into 5 subsets denoted by letters A to E. In our experiments the subsets A, B and C were used for training, subset D as development set and subset E as the test set. The Switchboard-1 Release 2 telephone speech corpus (LDC97S62) [15] has 2400 two-sided telephone conversations among 543 speakers from all over the United States. Switchboard-1 database contains continuous conversational speech which was recorded in real time. In our experiment the 33 hour subset of Switchboard-1 database in Kaldi toolkit recipe [16] was used.

5.2. Baseline Acoustic models

The baseline acoustic models were built with Mel frequency cepstral coefficients (MFCC) features. The recognition performance of baseline acoustic models are given in Table 3. The monophone model had 3 states for non-silence phones and 5 states for silence phones. For SVitchboard the triphone and linear discriminant analysis maximum likelihood linear transform (LDA-MLLT) models were built with 1300 context dependent states and 6 mixtures per state. The DNN was trained with 6 layers and 2048 nodes in each hidden layer. For Switchboard 33hr task the triphone and LDA-MLLT models were trained with 3200 context depend states and 10 mixtures per state. Switchboard 33hr task also used 6 layers and 2048 nodes for DNN training.

Now using each of the classifiers the articulatory features are extracted for SVitchboard database as shown in Figure 2. Acoustic models are built using the extracted features from each of the neural network architecture and the results are given in Table 2. Monophone and triphone models are built over AF with its delta and acceleration coefficients. LDA-MLLT model was built by splicing AF over 7 frames and projecting down to 40 dimension.

Table 3: Results in (%WER) for AF Extracted using AF Classifiers build with 110hr of Switch board using CTC criteria

Model Type	SVitchboard						Switchboard 33 hours		
	Test Set			Dev Set			Eval2000		
	MFCC	AF	AF-append	MFCC	AF	AF-append	MFCC	AF	AF-append
Monophone	70.92	55.25	52.68	70.39	55.07	52.96	69.9	64.8	39.9
Triphone	56.46	42.78	35.42	55.24	43.46	34.17	45	52.6	31.3
LDA-MLLT	53.46	32	32.11	52.21	32.54	31.31	41.1	30.2	28.7
DNN	44.66	28.63	**28.14**	43.53	29.03	**27.67**	31	**27**	27.2

5.3. Results for Comparison

The recognition performance for the acoustic models built using AF extracted with different articulatory classifiers built from different neural network architectures are given in Table 2. The results confirm that the AF are superior to MFCC features. The efficacy of the articulatory classifiers built with frame alignments were measured from the frame classification accuracy obtained in training and cross-validation set. It was found that AF-BLSTM and AF-CNN-DNN gave best results compared to other neural network architectures. AF-BLSTM had better classification accuracy for AF groups like Place and Vowel over AF-CNN-DNN. Since, AF groups Place and Vowel occupy the major portion of coefficients in AF, this resulted in AF-BLSTM giving superior performance. AF-BLSTM-CTC gives token accuracy which cannot be directly compared with frame classification accuracy but the performance of BLSTM-RNN and the CTC criterion makes the AF-BLSTM-CTC the best articulatory classifier. To summarize, from Table 2, it is clear that the among the various architectures using frame-level alignments, BLSTM gave the best recognition performance. Further, for the same BLSTM network using CTC gave significant improvements over using frame-level alignments.

6. Analysis of Articulatory Features

Since BLSTM-CTC classifiers generated the best articulatory features, in all subsequent analysis we will use these as the AF features. We investigate the efficacy of AF as stand alone features and also when appended with the conventional 39-dimension MFCC features (including delta and acceleration coefficients). For the second case where AF are appended with MFCC features, we apply PCA on the original 49-dimensional AF feature and reduce it to 40 dimension. The 40 dimensional AF are then appended with 39 dimensional MFCC features. The monophone and triphone models were built with these features and LDA-MLLT model was built with LDA features obtained by splicing the appended features over 3 frames and projecting down to 40 dimension. In both experiments the DNN models were built with the LDA features obtained from the corresponding articulatory features.

6.1. Results of Analysis

The results for the experiments with AF as stand alone and appended with MFCC for SVitchboard and Switchboard 33 hours task are given in Table 3. Results in Table 3 shows that AF perform better than MFCC features and when appended with MFCC features they give similar or better performance than when used alone. The AF gave an absolute improvement of 16% in SVitchboard and 4% in Switchboard 33 hours task in DNN acoustic model over LDA features. The AF seems to be giving better improvements in the case of low resource datasets, i.e. SVitchboard task.

6.2. AF in context of Speaker Normalization

The AF extracted using articulatory classifiers give a discrete set of feature values which are invariant across speakers. Hence, AF are robust to speaker variations. We investigate this property and compare it to conventional feature space maximum likelihood linear regression (fMLLR) normalization. From Table 4, it can be seen that AF alone provides performance comparable to fMLLR, indicating that they are inherently robust to speaker variations. Further a DNN model was built by appending the fMLLR features to the AF which yielded better performance. This indicates that both features carry some complementary information which improved the results when appended.

Table 4: Results of AF in Comparison with fMLLR Features for Switchboard 33 hours task

Model Type	LDA-MLLT	fMLLR	AF	fMLLR+AF
DNN	31.00	27.30	27.00	26.10

7. Conclusion

In this paper, a novel technique to build articulatory classifiers using CTC criterion was proposed. The proposed technique eliminates the requirement of frame-level alignments for training articulatory classifiers. The articulatory features extracted using the proposed technique gave a relative important of 36% for SVitchboard task and 13% for Switchboard 33 hours task over the conventional MFCC features. In this paper, a detailed comparison was performed between the proposed technique and the other methods that use frame level alignments. The comparison showed that articulatory classifiers built with BLSTM with CTC criterion gave relative improvement of 12% over the BLSTM with frame alignments.

8. Acknowledgements

We thank Simon King for providing us with the SVitchboard database for our experiments.

9. References

[1] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[2] O. Schmidbauer, “Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations,” in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 616–619.

[3] K. Elenius and G. Takács, “Phoneme recognition with an artificial neural network.” in *EUROSPEECH*, 1991.

[4] E. Eide, J. R. Rohlicek, H. Gish, and S. Mitter, “A linguistic feature representation of the speech waveform,” in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 483–486.

[5] L. Deng and D. Sun, “Phonetic classification and recognition using hmm representation of overlapping articulatory features for all classes of english sounds,” in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–45.

[6] K. Erler and G. H. Freeman, “An hmm-based speech recognizer using overlapping articulatory features,” *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2500–2513, 1996.

[7] K. Kirchhoff, G. A. Fink, and G. Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[8] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, “An articulatory feature-based tandem approach and factored observation modeling,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–645.

[9] J. Frankel, M. Magimai-doss, S. King, K. Livescu, and Ö. Çetin, “Articulatory feature classifiers trained on 2000 hours of telephone speech,” in *In Proc. Interspeech*, 2007.

[10] K. Livescu, A. Bezman, M. Borges, L. Yung, O. Cetin, J. Frankel, S. King, X. Xhi, L. Lavoie *et al.*, “Manual transcription of conversational speech at the articulatory feature level,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–953.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[12] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

[13] H. Sak, A. Senior, K. Rao, and F. Beaufays, “Fast and accurate recurrent neural network acoustic models for speech recognition,” *arXiv preprint arXiv:1507.06947*, 2015.

[14] S. King, C. Bartels, and J. Bilmesy, “SVitchboard 1: small vocabulary tasks from Switchboard,” in *Annual Conference of the International Speech Communication Association*, 2005, pp. 3385–3388.

[15] J. Godfrey and E. Holliman, “Switchboard-1 Release 2 LDC97S62,” *Linguistic Data Consortium*, 1993.

[16] 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 *Proc. ASRU*, December 2011.

[17] Y. Miao, M. Gowayyed, and F. Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” *arXiv preprint arXiv:1507.08240*, 2015.