Automatic Speech Recognition with Deep Neural Networks for Impaired Speech

Cristina España-Bonet[1,2] and José A. R. Fonollosa[1]

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Universität des Saarlandes, Saarbrücken, Germany
cristinae@cs.upc.edu, jose.fonollosa@upc.edu

Abstract. Automatic Speech Recognition has reached almost human performance in some controlled scenarios. However, recognition of impaired speech is a difficult task for two main reasons: data is (*i*) scarce and (*ii*) heterogeneous. In this work we train different architectures on a database of dysarthric speech. A comparison between architectures shows that, even with a small database, hybrid DNN-HMM models outperform classical GMM-HMM according to word error rate measures. A DNN is able to improve the recognition word error rate a 13% for subjects with dysarthria with respect to the best classical architecture. This improvement is higher than the one given by other deep neural networks such as CNNs, TDNNs and LSTMs. All the experiments have been done with the Kaldi toolkit for speech recognition for which we have adapted several recipes to deal with dysarthric speech and work on the TORGO database. These recipes are publicly available.

Keywords: speech recognition, speaker adaptation, deep learning, neural networks, dysarthria, Kaldi

1 Introduction

Automatic speech recognition (ASR) consists on automatically transcribing voice into text. It is not an easy task: one has do deal with noise, differences among speakers and spontaneous speech phenomena among others. For some controlled scenarios where one can minimise the effect of these phenomena, ASR approaches or exceeds the accuracy of humans on several benchmarks [2, 19].

Despite the good performance of the recently proposed end-to-end neural speech recognizers [2], Hidden Markov Models (HMM) are still the backbone of competitive speech recognition systems [19]. HMMs model speech signals with a sequence of states with an associated probability distribution for every observed vector. This probability can be represented using different approaches such as Gaussian mixture models (GMM) or artificial neural networks (ANN). In this

[0] A. Abad et al. (Eds.): IberSPEECH 2016, LNAI 10077, pp. 1–11, 2016.
DOI: 10.1007/978-3-319-49169-1_10

work, we refer to the former systems as *classical architectures* and to the latter ones as *neural network architectures*.

Although in its infancy ANNs were not able to deal with long time-sequences of speech signal by themselves, hybrid systems ANN-HMM already showed to be state-of-the-art at the beginning of the 90s [24]. ANNs solve at least two problems with respect to GMMs [3]: (*i*) assumptions about the shape of the statistical distribution of input features are not required and (*ii*) all training data is used to train a state (an not only that aligned to that state). On the opposite, they need of larger computing capabilities especially for large vocabularies.

Currently, and due to the existence of huge computing capabilities, hybrid deep neural network architectures DNN-HMM have been able to improve significantly ASR with respect to GMM-HMM systems for large vocabulary tasks [21, 4, 8]. Increasing the number of neurons and hidden layers in the network improves the word error rate (WER) in the recognition. However, for sparse data —small data sets— such amount of parameters cannot be properly fit and performance diminishes [10].

When dealing with impaired speech, one must face the problems of data sparsity. Since gathering data is even more difficult in this case, few databases exist, and the ones that exist are small. Besides, differences among speakers are larger and databases tend to be more heterogeneous. This poses a problem for ANNs, but also for GMMs which are more sensitive to differences between training and test data.

Here, we study the performance of both classical and neural network architectures when training on a small database of speakers with dysanthria, the TORGO database [9]. We discuss the differences not only between classical and neural systems, but also the suitability of using speaker adaptation techniques in this case. All the systems are trained using the Kaldi Speech Recognition toolkit [15]. We have adapted several recipes in order to prepare the data, extract the features and train the systems[3].

The remaining of the paper is organized as follows. First, Section 2 describes the database we use for the experiments. Next, Section 3 introduces the main architecture of an ASR and the specific techniques and resources we use. Section 4 makes emphasis in the acoustic model module and presents the different recognition systems that are evaluated in this task. Finally, we discuss the results and draw the conclusions in Sections 5 and 6 respectively.

2 The TORGO Database

Serveral speech disorders can alter the correct uttering of sounds. Speakers with dysarthria show difficulties to articulate phonemes due to a lesion in the nervous system. This may cause changes in voice quality, slow rate of speech or abnormal pitch and rhythm.

[3] Recipes are publicly available at `https://github.com/cristinae/ASRdys`

Table 1. Figures for the 15 speakers in the TORGO database ranked according to their degree of disorder

Speaker	F01	M01	M02	M04	M05	F03	F04	M03
Degree	severe	severe	severe	severe	sev-mid	mid	mild	mild
#audios	228	739	772	659	610	1097	675	806

Speaker	FC01	FC02	FC03	MC01	MC02	MC03	MC04
Degree	none	none	none	none	none	none	none
#audios	296	2183	1924	2141	1112	1661	1614

The TORGO database [9] contains speeches from 15 subjects, 6 females and 9 males. In total, the database contains about 3 hours per speaker of recorded speech, and one third corresponds to impaired speech. Four speakers have severe dysanthria, one is moderately-to-severely dysanthric and one is moderately dysanthric. Two other subjects have very mild dysanthria and the remaining 7 subjects are control speakers without any disorder. Table 1 describes the 15 subjects and includes the number of audios available in the database. For most of the utterances we use both, the audio obtained with a head-mounted microphone and the one obtained with a directional microphone.

Using the two microphones, we have 5586 utterances for speakers with dysanthria (a mean of 698 per speaker) and 10931 utterances for control speakers (a mean of 1562). An utterance can be a single word or a sentence, and the mean of words per utterance is of 3.5.

3 System's Architecture

All the systems described in the following sections share a common main architecture with four modules: (*i*) feature extraction, (*ii*) acoustic modeling, (*iii*) language modeling and (*iv*) pronunciation lexicon. Only feature extraction and acoustic modeling differ among systems.

3.1 Feature Extraction

As basic acoustic features we use 13 Mel-frequency cepstral coefficients (MFCCs). The features are generated in 25 ms windows shifted by 10 ms for the control speakers and 15 ms for dysanthric speech. This configuration for dysanthric speakers was shown to be adequate in Ref. [9]. As explained in Section 2, this disorder can make speakers talk slower and widening the shift between consecutive frames helps to homogenise the differences between patients and control speakers. For convolutional neural networks (CNN), we use 40 dimensional filterbank features in order to account for the correlations in the signal, estimated at the same window intervals.

Besides, in order to obtain more evolved speaker independent (SI) features, we apply a Linear Discriminative Analysis transformation (LDA) for projecting sequences of frames into 40 dimensions and, afterwards, a Maximum Likelihood Linear Transformation (MLLT) to diagonalise the matrix and gather the correlations among vectors [6]. For speaker dependent features (SD), we apply a feature-space Maximum Likelihood Linear Regression (fMLLR) [16]. In some cases, we also add 100-dimensional iVectors to gather specific information for every speaker and for the environment [5, 20].

3.2 Acoustic Modeling

In this work we use a monophone model and several standard three-state context dependent triphone models that differ on the features used, the training methodology and how the probability associated to each HMM state is calculated. Section 4 describes the main characteristics of the acoustic models used.

3.3 Language Modeling and Pronunciation Lexicon

The SRILM Toolkit [22] is used to build a standard 3-gram language model with interpolated Kneser-Ney discounting on the training data transcripts. For the lexicon, we choose the Carnegie Mellon University Pronouncing Dictionary[4] for North American English. It contains over 134,000 words and their pronunciations in the ARPAbet phoneme set with 39 phonemes.

4 Acoustic Modeling

Two types of models are distinguished in the following subsections: classical architectures GMM-HMM and hybrid neural network architectures DNN-HMM.

4.1 Classical Architectures

We study different variations on the nature of the features and the kind of training used in a standard GMM-HMM architecture. Below, we list the systems analysed in this work with their main characteristics. For an easy comparison, we also show for every system and between parentheses the nomenclature used in Kaldi. We have adapted Kaldi's recipes to fit our data, and trained 7 classical systems with 1800 HMM states and a total of 9000 Gaussians:

MONO Monophone model with MFCC features (*mono*)

TRI Basic triphone model with features MFCC+Δ+$\Delta\Delta$ (*tri2a*)

TRI-SI Triphone model with speaker-independent transformations applied MFCC+LDA+MLLT (*tri2b*)

[4] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

TRI-SD Triphone model with speaker-dependent transformations added MFCC+LDA+MLLT+fMLLR (*tri3b*)

TRI-SDdis Triphone model TRI-SD with a discriminative Maximum Mutual Information (MMI) and a feature-space MMI training (fMMI), TRI-SD+MMI+fMMI. We use a learning rate of 0.001 (*tri3b_fmmi*)

Several discriminative trainings can be done to fit the HMM parameters. We have done experiments with MMI training, boosted MMI, Minimum Phone Error (MPE), and direct and indirect feature-space discriminative MMI training (fMMI) with several learning rates. Model TRI-SDdis is the best performing one for dysarthric speakers and, therefore, it is the one included in the analysis.

Finally, we also consider subspace Gaussian Mixture Models [14] with 8000 states and 19000 substates:

sGMM Subspace GMM on top of SD features MFCC+LDA+MLLT+fMLLR (*sgmm2_4a*)

sGMM2 Subspace GMM with additional speaker adapted transformations fMLLR (*sgmm2_4a_fmllr*)

4.2 Neural Network Architectures

In hybrid systems, ANNs are trained to estimate the probabilities of the HMM states. Different networks and configurations can be used for this purpose:

DNN$_{\text{CE}}$ Deep Neural Network trained on alignments obtained with MFCC+ +LDA+MLLT+fMLLR features using cross-entropy. The DNN has 6 hidden layers, 1024 neurons and 1800 output units. The net is initialised with stacked restricted Boltzmann machines (RBMs) (*dnn4b_pretrain-dbn_dnn*)

DNN$_{\text{sMBR}}$ We introduce a sequence discriminative training that minimises the error on the state labels in a sentence. Departing from DNN$_{\text{CE}}$, 6 iterations of state-level minimum Bayes risk (sMBR) are applied (*dnn4b_pretrain-dbn_dnn_smbr*)

Notice that several kinds of sequence-discriminative training can be used. Reference [25] presents experiments with MMI, MPE, sMBR and boosted MMI. Although their training sets are larger (300h and 110h) only small differences were found among objective functions, being slightly better sMBR, the one we use in the following sections.

CNN$_{\text{ba}}$ CNN with convolution along the frequency axis. It uses 40-dim filterbank features, two convolutional layers and a learning rate of 0.008 (*cnn4c*)

CNN$_{\text{sMBR}}$ A DNN$_{\text{sMBR}}$ is built on top CNN$_{\text{ba}}$. First, the CNN is trained and then we build RBMs on top, train a 6-layer DNN with cross-entropy and afterwards 6 iterations of discriminative training (*cnn4c_pretrain-dbn_dnn_smbr*)

Finally, we select two kinds of neural networks especially devoted to deal with time sequences: time delay neural networks and recurrent neural networks.

Table 2. WER scores for the 8 speakers with dysarthria and a set of selected systems.

	F01	M01	M02	M04	M05	F03	F04	M03
MONO	70.86	80.10	76.55	88.62	77.71	57.02	29.10	43.32
TR1	70.68	91.18	81.09	88.62	84.59	41.80	18.62	26.01
TRI-SI	76.80	79.12	83.67	88.68	96.71	53.08	18.97	32.59
TRI-SD	47.30	78.91	68.49	81.16	97.16	42.88	13.29	17.06
TRI-SDdis	45.68	74.74	66.49	79.29	70.46	39.87	12.82	11.57
sGMM	43.71	77.83	64.01	71.46	98.43	37.26	11.42	10.19
sGMM2	43.53	78.37	63.33	71.34	97.31	37.22	11.24	**9.74**
DNN_{CE}	39.57	**62.20**	**42.89**	**69.05**	62.60	39.30	13.06	17.71
DNN_{sMBR}	**35.61**	62.30	47.95	69.30	**62.53**	37.01	**10.95**	12.76
CNNba	53.24	66.04	77.66	83.62	65.67	46.78	15.81	37.88
CNN_{sMBR}	53.06	66.74	50.47	81.40	65.74	**33.89**	11.24	10.44
TDNN	66.19	69.50	62.28	73.51	88.18	47.46	14.34	28.04
TDNNiV	94.96	95.62	84.14	92.59	93.94	91.98	39.29	70.97
LSTM	59.71	71.61	67.33	72.97	84.73	48.28	12.00	27.50
LSTMiV	71.04	75.01	76.13	77.30	72.85	69.33	19.61	32.20

TDNN Multi-splice Time Delay Neural Network trained on alignments obtained with MFCC+LDA+MLLT+fMLLR features. It uses high-resolution MFCC features. The network has 3 hidden layers with p-norm input dimension of 2000 and output dimension of 250. The learning rate evolves from 0.01 to 0.007 (*nnet_tdnn_a_noIvec*)

TDNNiV Same characteristics as the previous network but we add 100-dim iVectors to the 40-dim high-resolution MFCC input features for speaker adaptation (*nnet_tdnn_a*)

LSTM Long-Term Short-Term Memory network with 3 hidden layers with 1024 neurons. The network is trained for 10 epochs with a learning rate that evolves from 0.0012 and 0.00036, and with momentum 0.5 (*lstm_noIvec*)

LSTMiV Same characteristics as the previous network but we add 100-dim iVectors to the 40-dim MFCC input (*lstm_ivec*)

5 Results and Discussion

We use 14 speakers for training the parameters in the acoustic model and test the systems on the 15th. So, during training, there is no distinction between speakers with and without dysarthria besides the different shift in the frame definition for extracting the features. Since there is few data especially for dysarthric speakers, a training done only with impaired speech does not improve the results. Similarly, the language model used for testing is estimated on the same 14 speakers, and including additional corpora of a different domain to train the language model does not improve the results either. For cross-validation in neural networks training, we always use the data coming from a speaker with mild

Table 3. WER scores for the 7 control speakers.

	FC01	FC02	FC03	MC01	MC02	MC03	MC04
MONO	22.40	30.27	29.88	39.52	42.99	33.59	51.19
TR1	13.06	24.06	23.38	36.73	30.66	30.10	42.07
TRI-SI	13.20	24.73	26.30	38.30	32.98	33.28	42.48
TRI-SD	8.01	21.96	16.71	16.96	19.46	27.47	36.27
TRI-SDdis	7.42	21.72	15.43	17.49	18.23	26.51	37.40
sGMM	7.86	20.87	13.57	15.12	16.16	24.89	28.82
sGMM2	7.57	20.93	13.55	14.90	16.12	24.96	28.38
DNN_{CE}	6.53	19.62	11.01	15.32	14.72	22.11	27.06
DNN_{sMBR}	**6.38**	19.24	**10.41**	**12.03**	**13.38**	**20.37**	23.76
CNNba	15.58	22.38	14.63	25.01	38.25	33.25	44.66
CNN_{sMBR}	9.64	**18.28**	12.35	11.65	15.91	23.79	38.16
TDNN	10.98	18.97	12.90	35.67	58.69	32.90	31.90
TDNNiV	16.32	24.51	21.48	51.31	62.00	49.92	62.99
LSTM	8.46	19.30	13.21	24.06	41.80	25.53	**21.78**
LSTMiV	**6.38**	19.93	13.91	21.47	37.20	27.25	29.92

dysarthria regardless the nature of the test speaker —we use subject F03, or F04 in case the test subject is F03.

Table 2 shows the results for speakers with dysarthria. We measure the quality of the systems by means of WER. Notice that for severe dysarthric speakers, triphone models are not able to improve on monophone models. In fact, for these speakers, significant improvements in the WER score only appear when speaker dependent transformations are applied. The same happens for control speakers (Table 3) but in this case a base triphone system is always better than monophone systems. In general, intrinsic differences among the 15 speakers make necessary speaker adaptation techniques.

Subspace Gaussian Mixture Models are the best performing classical models. Only in cases where the recognition is extremely difficult (M01 and M05) the TRI-SDdis system outperforms the sGMM family. It is remarkable the hardness of the task: whereas the mean error rate for control speakers is an 18%, the mean for the six patients with the most severe disease reaches a 65%. For patients with a mild pathology WER is lower and equivalent to that of control speakers.

The best performing network resulted to be a DNN trained with GMM-HMM alignments. DNN-HMM systems show the lowest WERs for 11 out of 15 test speakers, 6 out of 8 for the speakers with dysarthria. If we consider all the neural network architectures compared to the classical ones, these figures increase to 14 out of 15 and 7 out of 8 respectively. For subjects with a severe disease, there is no difference between a DNN only trained by minimising cross-entropy (DNN_{CE}) and that including a subsequent sequence discriminative training (DNN_{sMBR}), the mean error rate varies from 52.6 to 52.5. For the control speakers, WER diminishes from 16.6 to 15.1 when adding the discriminative training.

Several works report improvements using CNNs, TDNNs and LSTMs with respect to DNNs, especially for large vocabularies [17, 1, 13, 18, 7, 11]. We do not find this behaviour in our task. The reasons are twofold: the TORGO database is small and data are heterogeneous. For comparison with other small databases, the authors in Ref. [13] train a TDNN on the Resource Management database, with about 3 hours of recorded speech. In their study, a standard DNN performed slightly better, although for larger amount of data a TDNN got better results. On the other hand, CNNs outperfom DNNs on the 50-hours English Broadcast News task [17] and on the 18-hours Microsoft-internal voice search task [1].

The neural network architectures we present apply speaker adaptation, at least through fMLLR features in the seed classical model and/or in the training of the network itself. For TDNNs and LSTMs, we also study the consequences of including iVectors. Although in other studies with larger databases the inclusion of iVectors improves a baseline without [12, 23], in our task it clearly damages the performance. For TDNNs the system with iVectors TDNNiV increments the WER in 24 points for dysarthric speakers and 12 points for control speakers. Results are not so negative for LSTMs but there is still a preference for the base LSTM: for dysarthric speakers the inclusion of iVectors causes an increment of 6 points of WER and for control speakers both systems are even.

This work is not the only one devoted to build an ASR for dysarthric speakers. The creators of the TORGO database trained an ASR in Ref. [9]. In their analysis, as in ours, simple triphone models are not able to improve significantly monophone models. So, instead of experimenting with new architectures built on triphone models, their approach is based on adapting speaker and acoustic models to incorporate a specific lexicon for each speaker. This lexicon includes pronunciations for several words that follow the guidelines of pronunciation detected in patients with dysarthria. When adapting the acoustic models to dysarthric speakers, the authors report a relative improvement in WER of 23% for the average of the 6 speakers with more severe dysarthria and a 3% further with the addition of the lexicon.

The creation of lexicons is difficult to generalize automatically since it depends on an analysis of the errors committed by each new speaker. Within our approach, we hope that deep neural networks can learn this behaviour from other speakers with similar problems. Our speaker adaptation models such as TRI-SD and sGMM2 are similar to that in [9]. Still, for the same 6 speakers, we get minor improvements with this adaptation: the sGMM2 model achieves and improvement in WER of a 10% with respect to the baseline, much smaller than their 23%. The difference is given mainly by the subject M05, while the best models in [9] reach a 15% WER, our models do not surpass a 70% WER for this speaker. The difference can only be explained by different data. Our best architecture with neural networks, $\mathrm{DNN}_{\mathrm{sMBR}}$, achieves a 23% of improvement in WER compared to baseline, which is similar to that in Ref. [9] but without building any resource manually.

6 Conclusions

Recognising dysarthric speech is a difficult task. We have trained an ASR for speakers with and without dysarthria using the TORGO database, a database of dysarthric articulation. With about three hours of recorded speech per each of the 15 subjects, moderate word error rate scores are obtained even for control speakers. A mean WER of 15% is obtained in this case, while it rises to 52% for the six test patients with more severe dysarthria.

Hybrid DNN-HMM systems are those with a best performance. DNNs outperfom the best classical system in 14 out of 15 test speakers: the WER score is improved a 3% for control speakers and a 13% for subjects with dysarthria with respect to the best classical architecture, a subspace GMM model with additional speaker adapted transformations fMLLR. Both in classical and neural architectures, speaker adaptation techniques are important for improving the recognition. For classical systems, fMLLR transformations make a qualitative leap with respect to the speaker independent transformations MLLT. Neural networks use TRI-SD models for training. However, in this task, iVector caracterisations of the speaker and the environment have a negative impact on the quality of the final systems.

Current results have been obtained using a database that combines impaired and normal speech. It remains to be seen whether including additional data for normal speech is able to further improve the recognition.

Acknowledgements

This work was supported by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund, contract INNPACTO IPT-2012-0914-300000 and TEC2015-69266-P (MINECO/FEDER, UE).

References

1. Abdel-Hamid, O., Deng, L., Yu, D.: Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Interspeech 2013, Lyon, France, August 25-29, 2013. pp. 3366–3370 (2013)
2. Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B.C., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A.Y., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A.Y., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., Zhu, Z.: Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. CoRR abs/1512.02595 (2015)
3. Bourlard, H.A., Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Norwell, MA, USA (1993)
4. Dahl, G., Yu, D., Deng, L., Acero, A.: Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 30–42 (January 2012)

5. Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Interspeech 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009. pp. 1559–1562 (2009)
6. Gopinath, R.A.: Maximum likelihood modeling with Gaussian distributions for classification. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998. pp. 661–664 (1998)
7. Li, X., Wu, X.: Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. pp. 4520–4524 (2015)
8. Maas, A.L., Hannun, A.Y., Lengerich, C.T., Qi, P., Jurafsky, D., Ng, A.Y.: Increasing Deep Neural Network Acoustic Model Size for Large Vocabulary Continuous Speech Recognition. CoRR abs/1406.7806 (2014)
9. Mengistu, K.T., Rudzicz, F.: Adapting acoustic and lexical models to dysarthric speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP11). pp. 4924–4927. IEEE (2011)
10. Miao, Y., Metze, F.: Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. In: Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier (eds.) Interspeech. pp. 2237–2241. ISCA (2013)
11. Miao, Y., Metze, F.: On speaker adaptation of long short-term memory recurrent neural networks. In: INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015. pp. 1101–1105 (2015)
12. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.: Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: Interspeech 2015, Dresden, Germany, September 6-10, 2015. pp. 2440–2444 (2015)
13. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Interspeech 2015, Dresden, Germany, September 6-10, 2015. pp. 3214–3218 (2015)
14. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R.C., Schwarz, P., Thomas, S.: The Subspace Gaussian Mixture model-A Structured Model for Speech Recognition. Comput. Speech Lang. 25(2), 404–439 (Apr 2011)
15. Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., Schwarz, P., Stemmer, G.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
16. Povey, D., Saon, G.: Feature and model space speaker adaptation with full covariance Gaussians. In: Interspeech 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006. ISCA (2006)
17. Sainath, T.N., Mohamed, A., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for LVCSR. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. pp. 8614–8618 (2013)
18. Sak, H., Senior, A.W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Interspeech 2014, Singapore, September 14-18, 2014. pp. 338–342 (2014)

19. Saon, G., Sercu, T., Rennie, S.J., Kuo, H.J.: The IBM 2016 English Conversational Telephone Speech Recognition System. CoRR abs/1604.08242 (2016)
20. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: ASRU. pp. 55–59. IEEE (2013)
21. Seide, F., Li, G., Yu, D.: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In: Interspeech 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011. pp. 437–440 (2011)
22. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of the Seventh International Conference of Spoken Language Processing (ICSLP2002). pp. 901–904. Denver, Colorado, USA (2002)
23. Tan, T., Qian, Y., Yu, D., Kundu, S., Lu, L., Sim, K.C., Xiao, X., Zhang, Y.: Speaker-aware training of LSTM-RNNS for acoustic modelling. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016. pp. 5280–5284 (2016)
24. Trentin, E., Gori, M.: A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing 37(14), 91–126 (2001)
25. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Interspeech 2013, Lyon, France, August 25-29, 2013. pp. 2345–2349 (2013)