IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program transcription

Juan M. Perero-Codosero[1,2], *Javier Antón-Martín*[1,2], *Daniel Tapias Merino*[1], *Eduardo López-Gonzalo*[2], *Luis A. Hernández-Gómez*[2]

[1]Sigma Technologies S.L.
[2]GAPS Signal Processing Applications Group, Universidad Politécnica de Madrid
{jmperero, janton, daniel}@sigma-ai.com, {eduardo.lopez, luisalfonso.hernandez}@upm.es

Abstract

Deep Neural Networks (DNN) are fundamental part of current ASR. State-of-the-art are hybrid models in which acoustic models (AM) are designed using neural networks. However, there is an increasing interest in developing end-to-end Deep Learning solutions where a neural network is trained to predict character/grapheme or sub-word sequences which can be converted directly to words. Though several promising results have been reported for end-to-end ASR systems, it is still not clear if they are capable to unseat hybrid systems.

In this contribution, we evaluate open-source state-of-the-art hybrid and end-to-end Deep Learning ASR under the IberSpeech-RTVE Speech to Text Transcription Challenge. The hybrid ASR is based on Kaldi and Wav2Letter will be the end-to-end framework. Experiments were carried out using 6 hours of dev1 and dev2 partitions. The lowest WER on the reference TV show (LM-20171107) was 22.23% for the hybrid system (lowercase format without punctuation). Major limitation for Wav2Letter has been a high training computational demand (between 6 hours and 1 day/epoch, depending on the training set). This forced us to stop the training process to meet the Challenge deadline. But we believe that with more training time it will provide competitive results with the hybrid system.

Index Terms: TV shows Speech-to-Text transcription, ASR systems, Hybrid DNN-HMM, End-to-end Deep Learning.

1. Introduction

Deep Neural Networks (DNN) have become fundamental part of current ASR systems. State-of-the-art approaches are generally hybrid models in which acoustic models (AM) are designed using neural networks to create HMM class posterior probabilities. These HMM-based neural network acoustic models (DNN-HMM) are combined with conventional pronunciation (PM) and language (LM) models [1].

The main limitations in hybrid ASR systems is a high complexity associated to the boot-strapping process for training the DNN-HMM models, requiring phoneme alignments for frame-wise cross entropy, and a sophisticated beam search decoder [2].

Though several approaches are being proposed to overcome these limitations, such as to train without requiring a phoneme alignment, or to avoid the lexicon [3], there is an increasing interest in working towards end-to-end solutions.

In end-to-end deep learning ASR systems [4] a neural network is trained to predict character/grapheme or sub-word sequences which can be converted directly to words, or even word sequences directly. They present the important advantage of integrating conventional separate acoustic, pronunciation and language models (AM, PM, LM) into a unified neural network modeling framework. Using a simplified training process acoustic, pronunciation and language modeling components are integrated to generate the hypothesized graphemes, sub-words or word sequences. This also greatly simplifies the decoding.

Most end-to-end ASR approaches [5] are typically based on a Connectionist Temporal Classification (CTC) framework [6, 7], a Sequence-to-Sequence attention-based encoder-decoder [8, 9] or a combination of both [10].

Several sequence-to-sequence Neural Network models have been proposed as Recurrent Neural Network Transducer (RNN-T) [11], Listen, Attend and Spell (LAS) [9], and Monotonic Alignments [12]. In attention-based encoder-decoder schemes [4] the listener encoder module plays a similar role that a conventional acoustic model, the attender learns alignments between the source and the target sequence, and the decoder works as a language model. Better performance has been reported [4] by modelling longer units such as word pieces models (WPM) and using Multi-head attention (MHA).

Several research works have also been proposed aiming to re-use HMM-based models to improve end-to-end systems. For example, well-trained tied-triphone acoustic models (AM) can be used as an initial model for a character-based end-to-end system [5] or training tied-triphone CTC models from scratch, but in this case a lexicon was required. However, these methods have important limitations as they demand a complex system development, high computation and a large amount of data for training, thus losing the attractiveness of end-to-end systems.

Therefore, even with the promising results already reported for many end-to-end ASR systems, it is still not clear if they are capable to unseat the current state-of-the-art hybrid DNN-HMM ASR systems.

In this paper, our aim is to contribute to the research towards the development of end-to-end ASR systems as an alternative to state-of-the-art hybrid ASR systems. For this purpose, we will develop and compare two open-source hybrid and end-to-end ASR systems for a specific speech-to-text task. As hybrid system the DNN-HMM Kaldi Toolkit [13] will be used, while Wav2Letter [14] will be the end-to-end framework. The speech-to-text task we will work on will be the RTVE IberSpeech 2018 Challenge[1]. This task represents a highly demanding domain corresponding to the automatic transcription of TV shows and broadcast news, in specific conditions, as different noisy environments and with the lack of accurate transcriptions for training.

The rest of the paper is structured as follows. In Section

[1]http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/

10.21437/IberSPEECH.2018-55

2, we describe the end-to-end and the hybrid ASR systems, which we have evaluated in the RTVE IberSpeech 2018 Challenge. Section 3 details the datasets we have used: how we have preprocessed them and the experimental protocols we have followed. Results are shown and discussed in Section 4. Finally, we summarize our results and conclusions in Section 5.

2. Deep Learning ASR Systems

2.1. End-to-end Speech Recognition System

As representative of open-source end-to-end ASR systems, we have chosen Wav2Letter. The Wav2Letter ASR system is based on a neural network architecture composed of convolutional units [15], with a Gated Linear Units (GLUs) implementation [16, 17]. The acoustic modeling is based on Mel-Frequency Spectral Coefficients (MFSC) which feed the Gated CNNs that generate letter scores at their outputs. These scores are processed by an alternative to CTC, the Auto Segmentation Criterion (ASG) leading to letter-based sequences (see Figure 1). In order to train the acoustic model, the feature extraction module computes 40-dimensional MFSCs, due to robustness to small time-warping distortions, as referred in [17].

Figure 1: *Architecture of the end-to-end ASR system based on Wav2Letter system. Adapted from [17].*

As commented before, the Neural Network architecture is trained to infer the segmentation of each letter in the training transcriptions using Auto Segmentation Criterion (ASG), an alternative criterion to Connectionist Temporal Classification (CTC). CTC takes into account all possible letter sequences, allowing a special blank state, which represents possible garbage frames between letters or the separation between repeated letters. In ASG blank states are replaced by the number of repetition of the previous letter, consequently a simpler graph is obtained [14]. Besides this graph that scores letter sequences depicting the right transcription, another graph is used to score of all letter sequences. Finally, a beam-search decoder (as described in [14]), is used at the last stage. It depends on a beam thresholding, histogram pruning and an optional language model.

2.2. Hybrid Speech Recognition System

In order to compare with end-to-end ASR, we have built a hybrid ASR system using open-source Kaldi Toolkit [13]. The ASR architecture consists of the classical sequence of three main modules: an acoustic model, a dictionary or pronunciation lexicon and a N-gram language model. These modules are combined for training and decoding using Weighted Finite-State Transducers (WFST) [18]. The acoustic modeling is based on Deep Neural Networks and Hidden Markov Models (DNN-HMM).

For the implementation of Kaldi DNN-HMM acoustic modelling we followed the so-called chain model [19], based on a subsampled time-delay neural network (TDNN) [20]. This implementation uses 3-fold reduced frame rate at the output of the network; this represents a significant reduction in decoding computation and the corresponding test time. The reduced frame rate requires HMM traversable in one transition; we use fixed transition probabilities in the HMM, and don't train them. Additionally, training DNN-HMM following a sequence-level objective function allowed its implementation as a maximum mutual information (MMI) criterion without lattices on GPU: doing a full forward-backward on a decoding graph derived from a phone n-gram language

Starting from the available transcript of the training speech data, training acoustic models is an iterative process of audio re-alignment starting from GMM/HMM monophone models and progressing to more accurate triphone models through re-training. For all of our experiments we used conventional feature pipe-line that involves splicing the 13-dimensional front-end MFCCs across 9 frames, followed by applying LDA to reduce the dimension to 40 and then further decorrelation using MLLT [21]. For initial GMM/HMM alignments speaker independent acoustic models were obtained using fMLLR. The input features to the neural network in DNN-HMM models was represented by a fixed transform that decorrelates a vector of (40*7)-dimensional features obtained by packing seven frames of 40-dimesional features MFCC(spliced) + LDA+MLLT+fMLLR corresponding to 3 frames on each side of the central frame.

To improve robustness mainly on speakers variability, speaker adaptive training (SAT) based on i-vectors was also implemented [22]. Speaker adaptive models were obtained by fine-tuning DNNs to a speaker-normalized feature space. On each frame a 100-dimensional i-vector is appended to the 40-dimensional acoustic space. In this extended acoustic space i-vectors may supply information about different sources of variability as speakers ID, so the network itself can do any feature normalization that is needed. To overcome some issues reported when test signals have substantially different energy levels than the training data, in our experiments the test-signal energies were energy-normalized to the average of the training data.

3. Experimental Setup

3.1. Datasets

3.1.1. RTVE2018 Database

In this evaluation, we investigate the performance of end-to-end and hybrid ASR systems on RTVE voice contents[2], a collection of TV shows and broadcast news from 2015 to 2018.

Training partition consists of audio files with subtitles, with the following limitations:

- Subtitles have been generated through a re-speaking procedure that sometimes summarizes what has been said, producing imprecise transcriptions.
- Transcriptions have not been supervised by humans.
- Timestamps are not properly aligned with the speech signal.

Trying to avoid the use of these low-quality transcriptions, which could cause confusion in the acoustic space, audio data was initially aligned by a baseline alignment system. This system was the same hybrid system described in Section 2.2. but trained using our own labeled databases, explained in Subsection 3.1.2. To improve the quality of these automatic transcriptions, they were undergone to a manual supervision process.

[2]http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

RTVE training partition consists of 460 hours, however due to our limitations in the manual supervision process, only two training datasets have been prepared for our experiments: RTVE_train350 (350 hours of train set) and RTVE_train100 (a RTVE_train350 subset of 100 hours). Validation datasets where extracted from the 10% training set for each RTVE_train350 and RTVE_train100 partitions. These validation datasets have been designed trying to cover the different scenarios in the whole RTVE training data: political and economic news, in-depth interviews, debate, live magazines, weather information, game and quiz shows. Consequently, for testing purposes, two development datasets have been defined as follows:

- RTVE_dev1: 5 hours have been selected in a balanced way in order to have an hour of each show type (e.g. 20H_dev1 is one hour of 20H program).
- RTVE_dev2: 1 hour has been selected carrying out the same procedure as RTVE_dev1.

3.1.2. Other Databases

Acoustic models have also been evaluated over *open* training condition, that is: by using additional datasets. To this end two additional datasets have been used to train the system.

- VESLIM: It consists of 103 hours of Spanish clean voice, where speakers read some sentences. More details in [23].
- OWNMEDIA: It contains 162 hours of TV programs, interviews, lectures and similar multimedia contents. This dataset contains manual transcriptions.

3.2. Training

From the hybrid ASR system, our AM was trained following a process based on the Switchboard Kaldi recipe (TDNN Chain models).

In order to create the PM for hybrid ASR system, it is used a set of 29 real phonemes (without silence phones). Instead, end-to-end system, the vocabulary contains 38 graphemes representing the standard Spanish alphabet plus stressed vowels, the apostrophe, symbols for repetitions and separation.

3.3. Resources

Experiments have been carried out using several computation resources. A server with 2 Xeon E5-2630v4, 2,2GHz, 10C/20TH and 3 GPUs Nvidia GTX 1080 Ti was used for hybrid ASR system. GPU for the DNN training and CPU for the HMM training and final decoding.

In order to train end-to-end models, more RAM was required, so it was used a GPU Nvidia Quadro P5000 (16 GB), for training and letter decoding.

4. Results

4.1. Hybrid ASR System

First, we compared the performance of our hybrid system increasing training data volume from 100 to 350 hours, and by adding additional training data from our external datasets.

Evaluation plans mentioned that a reference TV show (LM-20171107) has been used to obtain results with some commercial systems. This show is a live magazine covering Spanish current events and it has been used to obtain first results. As expected, more than 14% relative improvement in WER is obtained we adding all the available data.

Table I: *WER on reference TV (LM-20171107) show for acoustic models over closed and open training conditions (different train data sizes and language models).*

Hybrid systems	**WER(%)**
$RTVE\_train100 + LM\_subtitles$	26.01
$RTVE\_train350 + LM\_subtitles$	24.21
$RTVE\_train350 + LM\_supervised$	25.95
$RTVE\_train350 + LM\_subtsuperv$	23.47
$RTVE\_train350 + Others + LM\_subtsuperv$	23.20
$RTVE\_train350 + Others + LM\_open$	**22.23**

For testing on *closed* training condition, we used the full volume of RTVE training data that has been manually revised (RTVE_train350).

In addition to AM, LM has been trained with a different corpus. Four LMs were generated: LM_subtitles (based on subtitles given in RTVE database for the Challenge), LM_supervised (based on transcriptions of RTVE data training supervised by humans), LM_subtsuperv (based on two mentioned corpus) and LM_open (based on several corpus: news between 2015 and 2018, interviews, film captions and the two mentioned before).

As shown in Table I, adding supervised transcriptions from the training set to a subtitles-based LM, we achieved a WER of 23.47%, a 3% relative improvement over the same system using a language model trained only with subtitles. This improvement is mainly due to the fact that supervised transcriptions contains some conversational language features (i.e. false starts, truncated words, filler words, syntactic structure changes at talking time, etc.) that are generally omitted in subtitles because of respeaking procedure. In contrast, a LM only trained with supervised transcriptions did not provide better results. In this case, there were some tags in these transcriptions when words could not be confidently revised/transcribed (e.g. foreign names, mispronunciation, background noise, etc.). Inserting tags meant to include "unk" symbol to the LM and results were not as good as expected.

We next evaluated the systems over *open* condition, where we increased the amount of data for both AM and LM training. We combined RTVE training dataset and our own databases (see Section 3.1.2), resulting in a total over 600 hours of speech. As it can be seen in Table I the hybrid system provided a slight improvement. But, more importantly, WER went down to 22.23% when transcriptions from additional corpus were incorporated to train a LM. As a result, increasing both the amount of audio and transcription data will enable us to obtain the maximum performance, and to cover as much information as possible appearing in test files.

The division of development datasets according to the different show types makes it possible a deeper error analysis. Table II shows that models applied to TV programs as 20H and Millennium obtain the best results, a low WER of 14-17% for the best models. This could be explained because contents are daily news having good acoustic conditions (clean voice, only one speaker at time) and being better featured in LM. However, CA (Comando Actualidad) dataset contains some challenging scenarios (interviews at the street, background noise, overlapping, music). As a result, models achieve a high value of WER (49.51%).

Furthermore, it has to be emphasized that reference master of transcriptions was given without any review from our part. To evaluate the possible impact of transcription errors in the refer-

Table II: *WER(%) on the different datasets of models over a closed and open training conditions (different train data volume and language models). The duration of each dataset is an hour.*

	20H_dev1	AP_dev1	CA_dev1	LM_dev1	Mill_dev1	LN24H_dev1
Hybrid systems						
$RTVE\_train100 + LM\_subtitles$	17.67	24.36	51.95	25.41	19.59	27.47
$RTVE\_train350 + LM\_subtitles$	16.04	21.68	51.58	23.61	17.43	26.35
$RTVE\_train350 + LM\_supervised$	16.05	22.18	59.34	25.06	19.22	26.75
$RTVE\_train350 + LM\_subtsuperv$	15.38	21.43	49.67	22.86	17.81	25.15
$RTVE\_train350 + Others + LM\_subtsuperv$	15.21	21.95	**49.51**	22.30	18.51	25.09
$RTVE\_train350 + Others + LM\_open$	**14.88**	**20.94**	49.55	**21.44**	**17.01**	**24.13**
End-to-end system						
$RTVE\_train100$	71.61	75.38	87.06	72.35	69.49	76.70

ence master, we undergo a new test using an external dataset in which the reference master of transcriptions has been made manually and revised. This dataset consists of 3.5 hours of television news broadcasts (similar to 20H). Applying the same best models trained for RTVE IberSpeech 2018 Challenge, we obtained a WER of 8.51%, being the best results achieved so far.

4.2. End-to-end ASR System

Overall our results using Wav2Letter are still far from those of Kaldi-based hybrid ASR system. Nevertheless we must first acknowledge our limitations training Wav2Letter due to its high computational demand. This forced us to compare both systems using only our 100 hours training dataset: RTVE_train100. We obtained a WER on the same development dataset (TV shows of dev1 and dev2 partition) was 73.41% WER (on reference TV show). Applying LM, WER improved to 65.3%. Time consumption in decoding was 0.0088*RT (without applying LM) and it increased to 8.42*RT (applying LM). In addition, Table II also collects the rest of the results after evaluating end-to-end system over *closed* training condition.

We tried both to validate our Wav2Letter development and to analyze these results in the context of recent developments of Wav2Letter on read speech LibriSpeech in English [24]. To this end we compared evolution of the loss function per training epoch we obtained training with RTVE_train100, depicted in Figure 2, with some other developments using a similar amount of training data. In particular we developed Wav2Letter on 100 hours from LibriSpeech in English and from VESLIM Spanish dataset (see Section 3.1.2) which also contains about 100 hours of audio. As it can be seen analyzing the evolution of the loss functions in Figure 2, it seems that characters are better modeled applying clean speech in which people read sentences than applying contents in not controlled conditions, as in RTVE_train100. This limitation could be solved applying some type of data augmentation to introduce perturbed samples to model better the graphemes.

Last, we must also indicate that we tried to train end-to-end system over *open* training condition. However, when using all our training databases we found a high computational demand: using our resources (see Section 3.3) it took about 1 day/epoch (unlike training the system over *closed* condition 6.5 h/epoch) and the best results have been achieved after 40 epochs. For this reason, results have not been delivered to RTVE IberSpeech 2018 Challenge. In any case, our expectations for the end-to-end models are high, and therefore, we have considered to continue scientific research in this area: changing training strategies, applying optimization techniques, etc. in future works.

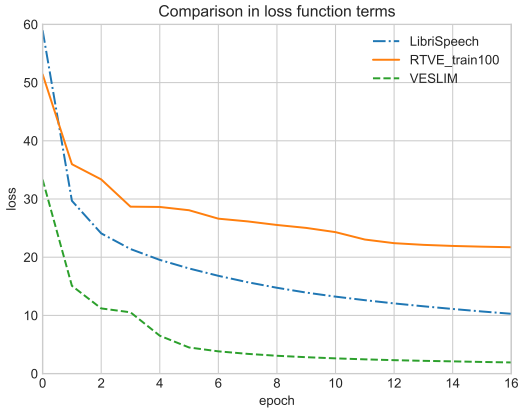

Figure 2: *Comparison in loss function terms during the end-to-end models training, on 100 hours of various databases: LibriSpeech, VESLIM and RTVE_train100.*

5. Conclusions

In this contribution we have tested the use of open-source hybrid and end-to-end ASR systems under the RTVE IberSpeech 2018 Challenge. According to our results, the development of a hybrid DNN-HMM Kaldi Toolkit [13] seems to be capable to address the difficulties of this hard task but it requires to put more effort in obtaining better quality transcriptions to improve both acoustic and language models. It is important to remark that WER of 8.51% is obtained, in the best conditions. However, further research on the use of robust feature spaces and DNN training should be addressed looking for robustness in the more challenging scenarios (street interviews with background noise, speakers overlapping, music, etc.). In what relates to end-to-end Wav2Letter system, we must acknowledge that our research has been limited by a high training computational time. Nevertheless, even under this limitation, we found that when compared to read speech (as LibriSpeech) it seems that both, the lack of correct transcriptions and the difficulties of dealing with conversational and noisy speech, will require more research so this kind of end-to-end architectures could be used for these challenging tasks.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep

neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv preprint arXiv:1805.03294*, 2018.

[3] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "Ctc in the context of generalized full-sum hmm training," in *Proc. Interspeech*, 2017, pp. 944–948.

[4] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[5] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," *arXiv preprint arXiv:1711.02212*, 2017.

[6] 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.

[7] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[9] 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.

[10] T. Hori, S. Watanabe, and J. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 518–529.

[11] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[12] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," *arXiv preprint arXiv:1704.00784*, 2017.

[13] 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.

[14] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv preprint arXiv:1609.03193*, 2016.

[15] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[16] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*, 2017, pp. 933–941.

[17] V. Liptchinsky, G. Synnaeve, and R. Collobert, "Letter-based speech recognition with gated convnets," *CoRR*, 2017.

[18] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[19] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[20] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

[22] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 1938–1949, 2015.

[23] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE transactions on speech and audio processing*, vol. 11, no. 6, pp. 617–625, 2003.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.