INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Recognition of Latin American Spanish using Multi-task Learning

Carlos Mendes[1,3], *Alberto Abad*[2,3], *João Paulo Neto*[1], *Isabel Trancoso*[2,3]

[1]VoiceInteraction, Portugal
[2]INESC-ID, Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Portugal

{carlos.mendes,joao.neto}@voiceinteraction.pt,{alberto.abad,isabel.trancoso}@inesc-id.pt

Abstract

In the broadcast news domain, national wide newscasters typically interact with communities with a diverse set of accents. One of the challenges in speech recognition is the performance degradation in the presence of these diverse conditions. Performance further aggravates when the accents are from other countries that share the same language. Extensive work has been conducted in this topic for languages such as English and Mandarin. Recently, TDNN based multi-task learning has received some attention in this area, with interesting results, typically using models trained with a variety of different accented corpora from a particular language. In this work, we look at the case of LATAM (Latin American) Spanish for its unique and distinctive accent variations. Because LATAM Spanish has historically been influenced by non-Spanish European migrations, we anticipated that LATAM based speech recognition performance can be further improved by including these influential languages, during a TDNN based multi-task training. Experiments show that including such languages in the training setup outperforms the single task acoustic model baseline. We also propose an automatic per-language weight selection strategy to regularize each language contribution during multi-task training.

Index Terms: Speech recognition, accented speech recognition, multi-task learning.

1. Introduction

Accents are one of the major challenges in speech recognition, and as such have received a wide attention from the speech research community, particularly for languages such as English and Mandarin [1, 2, 3, 4].

In the broadcast news domain, this problem is crucial, and yet there is relatively little work on accented speech recognition targeting this domain. We are particularly concerned with the study of cross-language influence. This topic has been widely studied in terms of language acquisition, and also in the scope of translation. For accented speech recognition, there are a few studies with non-native speech recognition[5, 6, 7], however, cross-language influence has not been so explored, to our knowledge. Yet, the possibility of using speech data from another language in the training seems worth investigating, in particular when there are much larger resources for this other language.

This work addresses this problem in the context of broadcast news recognition for LATAM (Latin American) Spanish, which has historically been influenced by non-Spanish European migrations, and in particular, for Argentinian Spanish.

Argentina is the largest Spanish-speaking country in the world, and within its borders there are several dialects influenced by distinct languages [8]. The vast majority of early European settlers in Argentina were from Spain, a country with a rich language diversity. A substantial Spanish descended Criollo population gradually built up in the new cities, while some mixed with the indigenous populations, with the black slave population, or with other immigrants from all over Europe. In fact, at some stage in history, some Argentinian dialects received more influence from Italian and Portuguese than other Spanish dialects. Therefore, Argentinian Spanish appears as an ideal language to investigate whether speech recognition can be improved by including influential languages in a multi-task training.

Multi-task learning is particularly well suited for hybrid DNN-HMM systems because it allows training a network with multiple target accents. Our objective is not to train a multi-language network or a low resource language, but rather to improve speech recognition for a specific accented language.

Our objectives are two folded: on one hand, we want to investigate whether one can improve accented speech recognition by incorporating multiple languages in the training process; and on the other hand we want to derive a strategy to automatically select per-language weights to regularize the contribution of each language during multi-task training.

This paper is structured as follows: Section 2 describes related work. The multitask learning process and the weight strategy are presented in Section 3. Section 4 describes the experimental broadcast news setup. Results are shown in Section 5. Finally conclusions and perspectives for future work are presented in Section 6.

2. Related Work

Accented speech recognition has been extensively studied in the past. Early traditional approaches to improve recognition of accented speech involved augmenting pronunciation lexica with accent-specific entries, namely learned from data [9]. Despite the significant reduction in cross-accent recognition error rates achieved by this type of approach, they present some limitations when acoustic models are tied to a specific accent, as models may not contemplate all possible allophonic variations.

Another common trend in accented speech recognition is the adaptation of acoustic models to multiples accents. Some of these techniques involve the adaptation of GMM-HMM-based models [3, 4] and DNN-based models [10].

More recent techniques use hierarchical grapheme-based end-to-end models [11]. However, they require large amounts of data to successfully recognize many accents.

Most recent developments in accented speech recognition uses multi-task learning[1] where a multi-task network is used in conjunction with accent embeddings. We take a different approach to multi-task learning, focusing on the study of accented languages with cross-language influences.

http://dx.doi.org/10.21437/Interspeech.2019-2772

3. Methodology

Our approach to accented speech recognition of LATAM Spanish is based on using multi-task learning for incorporating multiple languages/accents in the training process. This can be achieved by assigning each language and accent to a specific task.

3.1. Multitask learning

Multi-task [12] refers to the process of simultaneously learning multiple tasks from a single data set that contains annotations for different tasks. When applied to neural networks, a typical architecture consists of two distinct parts, the first part a sub-network shared by all tasks, and the second part a per-task specific output sub-network. During the training process, a loss function is back-propagated, for each task in turns, through both the task-specific sub-network and the shared sub-network. This type of architecture has been successfully applied in both single-language acoustic modelling [13, 14] and multi-lingual networks [15]. In this work, we use multi-task learning to train a network with multi-accent and multi-language tasks.

3.2. Dialect similarity and weighted training loss

One particular issue with multi-task learning is how to successfully weight each task during training. As an alternative to the typical approach of empirically calculating weights to balance the acoustic model training, we propose an automatic strategy to automatically regularize the training. This may be relevant to be able to include any language in the multi-task TDNN training, without concerns of biases or possible negative impacts on performance due to dissimilarities between the involved languages.

Our approach to automatically compute accent/language similarities is based on the representation of each accent/language as a single compact vector. This is achieved based on the i-vector or total variability space approach [16]. Total-variability modelling has emerged as one of the most powerful approaches to the problems of speaker and language verification. In this approach, the variability present in the high-dimensional GMM supervector is jointly modelled as a single low-rank total-variability space. The low-dimensionality total variability factors extracted from a given speech segment form a vector, named i-vector, which represents the speech segment in a very compact and efficient way. Thus, the total-variability modelling is used as a factor analysis based front-end extractor. The success of i-vector based speaker recognition has motivated the investigation of its application to other related fields, including language recognition tasks [17, 18, 19]. In this work, we extract per-segment i-vectors to compute a single average i-vector that represents each language. Modeling each language as a single vector is inspired by the work in [18], where the distribution of i-vectors for each language/accent are modelled with a single Gaussian. Then, given each accent/language average i-vector, the similarity between them is simply computed as the cosine similarity.

Two language i-vectors with the same orientation have a cosine similarity of 1; two orthogonal i-vectors have a similarity of 0, and two diametrically opposed i-vectors have a similarity of -1. Our strategy is based on using the scaled cosine similarity as weights, with values between 0 and 1, to regularize the frame-level cross-entropy loss for each language/accent. Given that this similarity is related to some distance between languages, it seems appropriate to use this information to weight the contribution of distinct languages, in the language independent part of the multi-task network. If a particular language has little similarity with the target language, then this means that its contribution to the network should be minimized.

Table 1: *List of languages and accents.*

LATAM Spanish Accents		**European Languages**	
es_AR	Argentina	es_ES	Spanish
es_CL	Chile	it_IT	Italian
es_CO	Colombia	de_DE	German
es_MX	Mexico		
es_PE	Peru		
es_PR	Puerto Rico		

4. Experimental Setup

As stated before, our goal is to improve speech recognition of Latin American Spanish dialects. However, given the vast range of Spanish dialects in America and the obvious experimental implications of analyzing as much as 22 Spanish varieties, we will focus on improving the particular case of Argentinean Spanish. The fact that Argentinean Spanish has historically been influenced by European migration [8] makes it a suitable candidate for analysis.

In this section, we describe our experimental setup. We begin by presenting the broadcast news corpus used in our experiments and the acoustic models generation process. Then, the Argentinian language model used in these experiments is presented, followed by a description of the i-vector cosine similarity models used in our analysis of language/accent proximity and in our proposed weighted loss regularization strategy.

4.1. Acoustic Model

Our experiments were conducted on 6 LATAM accents including Argentinian Spanish and 3 European Languages, see Table 1. The corpus used in this work was constructed by selecting approximately 30 hours per LATAM accent, from the VoiceInteraction in-house broadcast news corpora, and 130 hours per European language, from the SAVAS project corpora [20]. All dialect and language sets were further split into training, development and test sub-sets with ratios of 0.8, 0.1 and 0.1, respectively.

Acoustic models used in this work were built using the Kaldi toolkit [21] and trained using conventional Kaldi recipes. Language and dialect-specific senones and frame-level alignments were initialized with individually trained GMM-HMM tied-state triphone models. Each language/accent GMM-HMM system was trained in four stages: monophone flat-start; context-dependent triphones; context-dependent triphones on LDA+MLLT+fMLLR features[22]; and speaker adaptation training (SAT). The alignments from the speaker-adapted GMM-HMM systems were used to train the multi-task TDNN network with frame-level cross-entropy loss [13, 15, 23].

The multi-task network consists of two distinct parts: a first set of shared language/accent independent layers and a second set of language/accent-dependent output layers, one per each language/accent. The language/accent independent part of the network was configured with 6 TDNN hidden layers of 1024 units, spliced with offsets {0}, {-1,2}, {-3,3}, {-3,3}, {-7,2}, {0}, respectively. RELU activation functions were also used in the non-linear component of the TDNN hidden layers.

The input layer was configured to learn an affine transformation of the spliced frames over a window from t-2 to t+2. For each language/accent, a sub-network was appended to the language/accent independent layer. These sub-networks consist of a pre-final fully connected layer of 1024 units, with RELU activations, and a final softmax layer. The size of each output softmax layer is the number of probability density functions from the corresponding language/accent specific GMM-HMM models (ie. senones). The input to the network consists of a 26 dimension feature representation corresponding to 13 PLP coefficients plus deltas. Neither side speaker information (i-vectors), nor speed perturbation data augmentation have been considered in these experiments.

The optimization algorithm used to learn the network parameters was Natural Gradient for Stochastic Gradient Descent (NG-SGD) with exponential decaying learning rate [24]. All models were trained for 2 epochs with a minibatch size of 256.

4.2. Language Model

The language model used in this work is based on newspaper editions from Argentina and Spain publicly available on the web. The Argentinian database has a total of 66 million words and the Spain database has approximately 561 million words. We also used the Argentinian broadcast news transcriptions train set, from the VoiceInteraction in-house LATAM database, to model spontaneous speech. The total amount of words in transcriptions is 277 thousand. The Argentinian language model was trained using the SRI language modeling toolkit [25].

Because of the difference in terms of size, temporal and geographic contexts, we built two different back-off 4-gram language models for each country's newspapers sources, and a third back-off 3-gram language model for the Argentinian broadcast news transcriptions. Model perplexity was calculated using the Argentinian broadcast news transcriptions development set. The language model made from the Argentinian transcriptions yielded a perplexity of 320.1, and the models from the web newspapers obtained a perplexity of 319.3 (Argentinian newspapers) and 501.6 (Spanish newspapers).

The final language model was generated by interpolating all previous models with weight factors 0.47 (Argentinian newspapers), 0.12 (Spanish newspapers) and 0.41 (Argentinean transcriptions), optimized by minimizing perplexity in the Argentinian broadcast news transcriptions development set. The perplexity of the final language model scored 217.1 on the evaluation set.

In our recognition experiments we used the decoder from our in-house ASR system named AUDIMUS [26, 27]. The AUDIMUS decoder is based on the weighted finite-state transducer (WFST) approach to very large vocabulary speech recognition [28].

4.3. Dialect Similarity

The i-vector model used in this work was trained with data selected from the SAVAS project corpus[20] and the VoiceInteraction in-house LATAM corpus. The total amount of audio used for training was 2 hours per-language.

We followed the work from [17] and used the Bob toolkit [29] to train language-based i-vectors. Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [30] channel compensation techniques were used during training. The final model relies on a 1024 Gaussian UBM and a 200-dimension total variability matrix T.

Table 2: *WER with uniform weighted loss.*

Experiments	WER (%)	Relative Improvement (%)
es_AR, es_ES	12.79	7.56
es_AR, it_IT	12.96	6.37
es_AR, de_DE	13.85	-0.09
es_AR, es_ES, de_DE	12.91	6.67
es_AR, es_ES, it_IT	12.46	9.99
es_AR, it_IT, de_DE	12.98	6.18
es_AR, es_ES, it_IT, de_DE	12.30	11.11

This model was used to extract i-vectors for 100 speech segments per-language/accent. Then, the mean i-vector for each accent/language is calculated, which is used as a sort of accent/language single vector characterization. Notice that simple single Gaussian models of i-vectors trained with a relatively small number of i-vectors per language have proven to be very effective for language recognition tasks [18, 19]. Finally, the similarity of Argentinian to the other accents and languages is calculated as the cosine similarity among the corresponding mean i-vectors.

5. Results

In this section, we begin by presenting WER results obtained with the acoustic models trained with a uniform weighted loss function, followed by results for the i-vector cosine similarities for the Argentinian Spanish. We then report recognition results, when using our weight selection strategy during training, and finally we show how the Argentinian system can be further improved by combining LATAM and European languages in multi-task training.

5.1. Uniform weighted loss

Following the same procedure described in 4.1, we first trained a baseline TDNN acoustic model using only the Argentinian Spanish data. The WER obtained with this model, on the Argentinian Spanish test set, was 13.84%.

In the first set of experiments, we aimed at evaluating the impact of using training data from European languages for building the Argentinian models. For these experiments, we used Argentinian Spanish as target language, and European Spanish, Italian and German as complementary languages. We trained 7 different combinations of multi-task acoustic models with unitary and uniform weight loss. Thus, we expect to observe ASR improvements, specially when using data from the languages closer to the target one, i.e. with European Spanish and Italian.

In Table 2, results for the first set of experiments are presented. In the experiments where Argentinian was trained in conjunction with a single additional European Language, we can observe that the language with the most impact in terms of WER is European Spanish with 7.56% relative improvement over the baseline. Italian comes next, with 6.37%. The addition of German had no significant impact. In fact, the only combination where German seems to have any positive impact is when it is used with all other European languages, in which we report a relative improvement of 11.11% over the baseline.

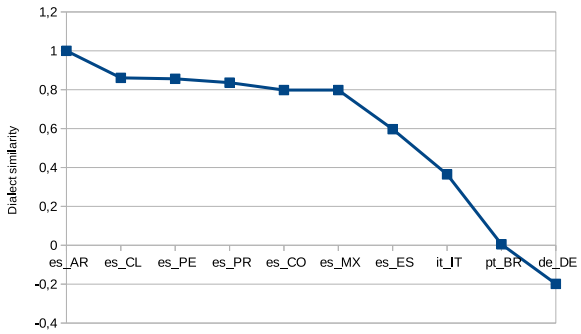

Figure 1: *Argentinian dialect similarities.*

Table 3: *WER with dialect similarity weighted loss.*

Experiments	WER (%)	Relative Improvement (%)
es_AR, es_ES	12.77	7.75
es_AR, it_IT	12.90	6.74
es_AR, de_DE	13.36	3.43
es_AR, es_ES, de_DE	12.65	8.61
es_AR, es_ES, it_IT	12.36	10.71
es_AR, it_IT, de_DE	12.89	6.88
es_AR, es_ES, it_IT, de_DE	12.34	10.81

5.2. Dialect similarity

Figure 1 shows the dialect similarities obtained for Argentinian Spanish with respect to the other LATAM Spanish accents and European languages. At a first glance, we can see how similar Argentinian Spanish is to the other LATAM dialects. On the other hand, of the Spanish dialects, European Spanish is the furthest from Argentinian Spanish. This result is consistent with the historical divergence between both countries, and the influences Argentinian Spanish received from other languages [8].

On the opposite side, German stands as the most distant language to Argentinian Spanish according to computed similarities. This is of course expected. German has been included in our experiments as a *control* language to study how the multi-task training behaves when languages that are very different to the target language are included in the training set.

5.3. Dialect similarity based weighted loss

In our second set of experiments, we sought to investigate the impact of our automatic per-language weight selection strategy on the Argentinian Models. We repeated the set of experiments in Section 5.1, but now with dialect similarity weights, normalized between 0 and 1, to regularize each language contribution during the multi-task training. We expect to observe ASR improvements, specially in the experiments using German training data.

In Table 3, results for the second set of experiments are presented. When comparing these results with results from Table 2, we observe that in general there are improvements over the previous set of experiments. The most significant results are the ones involving German. When used as the solo complementary language, we see relative improvements of 3.44% over the baseline, whereas in 2 no improvement is observed.

Table 4: *WER with combined accent and languages.*

Experiments	Weight Strategy	WER (%)	Relative Improvement (%)
LATAM	Uniform	12.40	10.41
LATAM + EU	Uniform	12.06	12.84
LATAM	Weighted	12.45	10.06
LATAM + EU	Weighted	11.99	13.33

We should note that, the only case where no improvement is registered is the experiment with all languages. In 2, we report a relative improvement of 11.11% and in this experiment we report an improvement of 10.81% over the baseline. This difference is not particular significant, because the WER diference between them is 0.04.

5.4. Multi-accent and multi-language combinination

In our last experiment, we explore the contribution of the European languages in a LATAM Spanish training scenario. The objective is to show that combining multi-accent sources with multi-language sources further improves the ASR results.

In Table 4, we present two experiments. The first a LATAM only multi-task training setup and the second a LATAM plus European languages multi-task training setup. In both experiments, we used dialect similarity weights. On the LATAM only setup we report relative improvements of 10.06% with respect to the baseline, and in the LATAM plus European setup we report relative improvements of 13.33% over the baseline.

Overall, the WER obtained with our final model, on the Argentinian Spanish test set, is 11.99%. This is a remarkable improvement with respect to the 13.84% WER achieved with the baseline system. This improvement has been obtained by simply using additional training data in a multi-task acoustic model architecture and by combining these additional sources in an informative way.

6. Conclusions

In this work, we explore a multi-task learning architecture to improve accented speech recognition, by incorporating multiple languages in the training process, using a data-driven strategy to automatically select per-language loss weights. The combination of both techniques is particular useful in accented languages with cross-language influences. We demonstrate this for the particular case of Argentinian Spanish, where our approach achieves a 13.33% relative improvement. Future work will extend the study to other accented languages, like e.g. Brazilian Portuguese, and the further exploration of the relation between the dialect similarity weights and empirical experiments on the training loss weights. We also intent to compare our weigth loss strategy with other techniques like multitask learning with an adversarial loss.

7. Acknowledgements

The authors would like to acknowledge our VoiceInteraction colleagues, in particular Tiago Luís and Alícia Martinez-Losa, for their contributions to this work. This work was partially supported by Portuguese national funds through *Fundação para a Ciência e a Tecnologia* (FCT) (reference UID/CEC/50021/2019).

8. References

[1] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep 2018, pp. 2454–2458.

[2] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2018.

[3] Y. Liu and P. Fung, "Multi-accent chinese speech recognition." in *INTERSPEECH 2006 – 9th Annual Conference of the International Speech Communication Association*, Pittsburgh, PA, USA, Sep 2006.

[4] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented english data," in *INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, Sep 2010.

[5] T. Tan and L. Besacier, "Acoustic model interpolation for non-native speech recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, April 2007, pp. IV–1009–IV–1012.

[6] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *INTERSPEECH 2006 – 7th Annual Conference of the International Speech Communication Association*, Pittsburgh, PA, USA, Sep 2006.

[7] S. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura, "Non-native english speech recognition using bilingual english lexicon and acoustic models," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, vol. 3, July 2003, pp. III–625.

[8] J. M. Lipski, *El Español de América*, 9th ed. Cátedra, 2017.

[9] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *ICSLP'96 - Fourth International Conference on Spoken Language Processing*, vol. 4, Banff, Canada, Oct. 1996, pp. 2324–2327.

[10] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2015.

[11] K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4815–4819.

[12] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[13] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6965–6969.

[14] P. Bell, P. Swietojanski, and S. Renals, "Multitask learning of context-dependent targets in deep neural network acoustic models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 238–247, Feb. 2017.

[15] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7304–7308.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[17] N. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, Aug 2011.

[18] D. M. González, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, Aug 2011.

[19] A. Abad, E. Ribeiro, F. Kepler, R. F. Astudillo, and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, San Francisco, CA, USA, Sep 2016.

[20] A. D. Pozo, C. Aliprandi, A. lvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, and M. Raffaelli, "SAVAS: Collecting, annotating and sharing audiovisual language resources for automatic subtitling," in *LREC'14 - Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 2014.

[21] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[22] S. P. Rath, D. Povey, K. Vesel, and J. ernock, "Improved feature processing for deep neural networks," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug 2013, pp. 109–113.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep 2015, pp. 3214–3218.

[24] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," in *Proc. ICLR workshop*, 2015.

[25] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *INTERSPEECH 2002 – 3rd Annual Conference of the International Speech Communication Association*, Denver, CO, USA, Sep 2002.

[26] H. Meinedo, N. Souto, and J. P. Neto, "Speech recognition of broadcast news for the European Portuguese language," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'01)*, Dec. 2001, pp. 319–322.

[27] H. Meinedo, A. Abad, T. Pellegrini, J. P. Neto, and I. Trancoso, "The L2F broadcast news speech recognition system," in *Fala2010*, 2010.

[28] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[29] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: A free signal processing and machine learning toolbox for researchers," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12. New York, NY, USA: ACM, 2012, pp. 1449–1452. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396517

[30] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *INTERSPEECH 2006 – 7th Annual Conference of the International Speech Communication Association*, Pittsburgh, PA, USA, Sep 2006.